A CRITERION-REFERENCED APPROACH TO TEST CONSTRUCTION IN PHYSICAL EDUCATION

by

SURINDER BRAR

B.P.E., The University of British Columbia, 1979

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTERS OF PHYSICAL EDUCATION

in

THE FACULTY OF GRADUATE STUDIES

(PHYSICAL EDUCATION)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

OCTOBER 1986

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Surinder Brar

Department of Physical Education

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date October 10, 1986

Abstract

The primary purpose of this study was to apply criterion-referenced (CR) test construction procedures in the development of a Basic Fitness Theory Exam (BFTE) for the British Columbia Recreation Association-Fitness Branch as part of a registration model for Basic Fitness Leaders. The BFTE will be used to determine whether or not individuals have sufficient knowledge to be certified as fitness instructors in British Columbia. A secondary purpose was to provide a clear set of procedures for CR test development of Physical Education knowledge tests.

The development of the BFTE consisted of a number of stages and involved three pilot tests. Pilot #1 consisted of 57 items and was administered to physical education students at The University of British Columbia (92 in Year I and 72 in Year II). After the results of the statistical analyses were presented to the Provincial Fitness Advisory Committee, 22 items were revised, 22 items were replaced, no items were deleted, 3 items were added, and 13 items were left unchanged. Pilot #2, consisting of 60 items, was administered to 94 instructed (upon completion of a 40 hour course) and 106 uninstructed (upon commencement of a 40 hour course) subjects. Again, comprehensive statistical analyses were performed and then the results were presented to the PFAC. In all, 15 items were deleted, one item was revised, and no items were added in constructing Pilot #3.

Pilot #3, the final test, consists of 45 items and is presently being used in the British Columbia Recreation Association-Fitness Branch's registration model for the Basic

Fitness Leader (Level I). The BFTE has a "pass" criterion of 70% (32 out of 45).

Table of Contents

List of Tables

List of Appendices

Introduction

The need for criterion-referenced (CR) tests was first demonstrated by Glaser (1963) over twenty years ago and subsequently expanded upon by Popham and Husek (1969). The purpose of the CR tests was to provide information relative to a domain of well-defined objectives or competencies which could then be used to make decisions regarding individuals (promotion, certification) or specific populations (program evaluation). Since the first discussions about CR testing there has been a growing trend in physical education to formalize competency evaluation. While physical educators have realized that motor skills as well as knowledge can be expicitly defined, specific levels of competency have only been defined in relatively few domains (e.g., British Columbia Ministry of Education, 1979, 1980; Shiflett & Schuman, 1982).

The British Columbia Physical Education Curriculum Guide (B.C. Ministry of Education, 1980) is an example of the effort made by physical educators to define levels of individual competency in a variety of activities. A list of specific behavioral objectives is given for each of the hierarchical levels of the activities, with the amount of detail given making it easy to use the objectives as a criterion for promotion of the students. Another example of individual competency evaluation and certification is the Canadian Association of Sport Sciences' certification and accreditation program for fitness appraiser's. A final example is the Physical Education

Learning Assessment Report (B.C. Ministry of Education, 1979) which shows how CR tests can be used to make evaluative statements about a specific population. Many other examples of both individual and programmatic decisions are available.

Due to the nature of the decisions which are made, based on CR testing, it is important that they be developed utilizing test construction procedures that are appropriate to their function. CR tests are, "those which are used to ascertain an individual's status with respect to some criterion" (Popham & Husek, 1969, p. 2), and show how an examinee stands with respect to some specific objective or domain of objectives. This is different from traditional or norm-referenced (NR) tests, "which are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device" (Popham & Husek, 1969, p. 2). NR tests provide little information regarding an individual's absolute degree of skill or competency, but rather provide the position of an individual relative to some sample or population (e.g., z-scores, percentiles).

Tests in physical education, including those mentioned above, are being used to make criterion-related decisions even though they were not developed according to CR test construction procedures. For example, Wilson (1980), developed a knowledge test according to standard NR methodology and when the test was administered the scores were interpreted as CR test scores. That is, the scores were interpreted in terms of a criterion that was not established during the test construction. Another area in which CR tests are being used in physical education is

in the evaluation of physical skills. For example, Shifflett & Schuman (1982) have developed a CR test for archery using a combination of CR and NR techniques. As with Wilson's knowledge exam, the interpretation of the archery test scores in a truely CR sense is hard to justify. The inappropriate interpretations shown above can result in organizations being accredited or individuals being certified or promoted because of high relative scores, but not necessarily high absolute scores. It is, therefore, necessary to develop, explain, and justify explicit test construction procedures for CR assessment in physical education.

Thus, one of the purposes of this study is to provide evidence, theoretical and empirical, of the procedures needed to develop a CR knowledge test in physical education. More specifically, the purpose of this study is to apply CR test construction procedures in the development of a Basic Fitness Theory Exam (BFTE) for the British Columbia Recreation Association-Fitness Branch as part of a registration model for Basic Fitness Leaders. The requirements of the model state that, in order to be eligible for registration, the individual must, "successfully complete a course which is recognized by the British Columbia Recreation Association, practical requirements, and Basic Fitness Theory Exam." The BFTE will be used to determine whether or not individuals have sufficient knowledge to be certified as fitness instructors.

Many authors have argued over the appropriateness of various psychometric indices and procedures (e.g., Berk, 1980a; Hambleton, Swaminathan, & Algina, 1978; Huynh, 1979; Subkoviak,

1980). Therefore, while developing the exam, this study will also contrast specific CR test construction procedures. Strengths, weaknesses, and applicability of the procedures to this specific situation will be discussed. Upon completion, this study will provide an exemplary model of test construction for CR tests in physical education.

Criterion-Referenced Test Construction: An Overview

Since CR testing was first popularized by Glaser (1963) and Popham and Husek (1969) many researchers have become involved in this rapidly growing area of study. Even as early as 1978, Hambleton et al. (1978) stated that there were over six hundred references on the topic of CR testing and that, "there are as many ideas about what a criterion-referenced test is as there are contributors to the field." The number of contributors has continued to increase over the past few years, however, much more consensus has been reached over definitions, and appropriateness of various CR test construction procedures. An overview of the developmental stages for constructing a CR test are shown in the first column of Table 1 below (adapted from Berk, 1980b, p. 6). For comparison, the second column of Table 1 shows the developmental stages for a traditional NR test. The following sections discuss the development of the different stages of CR test construction.

Item Construction

Table 1 shows that the first three stages for CR test construction are similar to the first four stages in NR test construction. However, CR test construction places more emphasis on clear objective and domain specification and on item generation procedures. CR test procedures could be used to generate items for a NR test, however, NR test construction procedures are not strict enough to generate good CR test items (Martuza, 1977, chap. 16).

Table 1

Test Construction Developmental Stages

Criterion-referenced	Norm-referenced
Content domain specification	Content outline
Item construction	Objective outline (LO)
Item domain construction	Content objective matrix
Item analysis	Write items/ construct test
Item selection	Item analysis
Cut-off score selection	Item selection
Validity	Validity
Reliability	Reliability
Score interpretations	Score interpretations

Hambleton et al. (1978, p.3) also confirm this by stating that, "a norm-referenced test can be used to make criterion-referenced measurement, and a criterion-referenced test can be used to make norm-referenced measurement, but neither use will be particularly satisfactory." It should be pointed out that CR test construction methodology has only developed over the past twenty years; until very recently CR tests were constructed using NR methodology and then the resulting scores were just interpreted in a CR sense. As mentioned above, this was a very unsatisfactory method of obtaining the information needed to make CR decisions.

The first stage in CR test construction is to define the domain of content or behaviors to be measured. This is a very important stage because the, "validity and interpretation of the test scores are contingent upon the precision of the domain specifications" (Berk, 1980b, p. 13). The precision used in this very first stage alone sets CR testing apart from standard NR testing where the content and objective outlines can be done quite arbitrarily. While most authors agree that this is a important stage, very few (Berk, 1980b; Hambleton et al. 1978; Millman, 1974) have actually discussed methodology for developing good domain specifications.

The next stages involve item and item domain construction. The items should be generated according to a specific set of rules to ensure that thay address the content defined in the first stage. A sample of item construction guidelines can be seen in Appendix A. The content domain and the item generation rules should be specified in such a rigid fashion that the items

produced are totally independent of the person writing them. Popham (in Berk, 1980b, chap.2) even discusses the use of algorithms for computer generated test items which would all be very uniform in nature and also very closely related to the domain specifications.

The item domain is constructed as a result of generating items within the content domain specified in the first stage. It can be either finite or infinite depending on the precision of the content domain specification strategy employed. A finite set of items will result from a very precisely defined content domain.

Item Analysis and Selection

Item analysis, and consequently item selection and revision, are based on item-objective congruence and various item statistics as listed in Appendix B. The data for these statistics are obtained by the following methods:

a) pretest-posttest

b) uninstructed-instructed group difference

c) individual gain

d) net gain

Although the methods for data collection are traditional, the actual statistics used are different, or in some cases just interpreted differently, than the standard NR statistics. The main reason for needing new statistics, or different interpretations, is due to the lack of variance in the CR test scores.

The first and most important analysis at the item level is

the item-objective congruence. It shows, "the extent to which an item measures the objective it is intended to measure" (Berk, 1980b, p.51). Each of the items should be subjectively rated by a series of individuals who are experts in the content domain specified in stage one. A summary of the judges ratings will show how well each item is addressing the objective it was meant to test. Ensuring a high level of item-objective congruence will help towards constructing a test with high content validity. A sample item-objective congruence rating sheet can be seen in Appendix C.

Informal examinee feedback. Feedback can be obtained by attatching a questionnaire to the test as it is administered. The questionnaire should consist of direct questions regarding ambiguities, confusing words, poor wording, difficulty level, etc. A list of possible questions can be found in Appendix D. This feedback can provide "valuable insights and directions for improving the test that otherwise would not be disclosed from a quantitative analysis" (Berk, 1980b).

Item difficulty. Item difficulty, expressed as a "p-value", refers to the percentage of examinees choosing the correct response. A higher p-value for the correct response indicates that many examinees answered correctly; the item may be too easy. It should be noted that this paper will refer to the p-value of the distractors as well as the correct response. In this case, the p-value for a given response will represent the percentage of examinees who chose that response, regardless of whether or not it is the correct response. The p-value of distractors can provide invaluable information regarding the

discrimination ability of the item.

Item discrimination. Item discrimination indices should demonstrate changes from pretest to posttest or differences between instructed and uninstructed subjects. These indices need to show maximum discrimination between groups and minimum discrimination among individuals within any one group. Berk (1980b) lists, describes, and evaluates several indices that have been suggested by various authors. These indices (shown in Appendix B) range from three simple proportions, which are very sensitive to the accurate definition of criterion groups, to three types of correlations which require more sophisticated computer programs for their calculations. The final index Berk (1980b) evaluates is Millman's (1974) discrimination index which involves a stepwise regression which incorporates interitem correlations. This index not only requires a sophisticated computer program, but also requires a very large sample size. Regardless of which of the indices is chosen, it is important to ensure that all items discriminate well; good item discrimination will result in a test with a high level of decision validity.

Item homogeneity. Item homogeneity refers to statistics that are used to verify, "that items congruent with an instructional objective behave similarily on a single testing or on repeated testings" (Berk, 1980b, p. 64). Item homogeneity is more appropriate in situations where the objective is very specifically defined; in a case where the objective is generally defined to cover a variety of skill it may be unrealistic to expect the items to be homogeneous in terms of item difficulty

indices, etc. Berk (1980b) evaluates four possible indices that can be used to demonstrate item homogeneity.

In general, item statistics should not serve as the sole criterion for item selection or revision, but rather should provide guidelines for these procedures. Millman (1974, p. 339) suggests that, "Item statistics can, however, be used to detect flawed items."

Cut-off Score Selection

Many researchers have attempted to solve the problem of determining the best method for establishing a cut-off score for criterion-referenced tests which would assign examinees to mastery/nonmastery groups (Hambleton & Eignor, 1979; Hambleton et al., 1978; Millman, 1973; Shepard, 1980). The problem has yet to be resolved to a point which is satisfactory. Hambleton et al (1978, p. 26) state that, "The arbitrariness of the proposed solution has proved troubling to some measurement people, to the point where they seriously question the merits of determining and using cut-off scores at all." Cut-off scores can also have a profound effect on the test score reliability and validity. Researchers have provided evidence which suggests that the value of a cut-off score influences student learning and their attitudes. A higher cut-off score corresponds to students who study harder and have a more positive attitude towards the test. As a result, their test scores will be higher.

The selection of cut-off scores will always be associated with errors. Examinees who are very close to the cut-off score

(just above or below) will be very similar in their abilities. Hambleton (1978), Shepard (1980), and others, while realizing that the standard setting methods are somewhat arbitrary, suggest that, "potentially flawed standards are better than none." If pass-fail decisions are inevitable, good test information, even with an arbitrary cut-off score, will lead to better decisions than those that would be made without the test.

Shepard (1980) reviews, evaluates, and summarizes exemplary standard-setting methods. The major methods fall in the following categories:

(1) absolute judgements of test content

(2) judgements about mastery-nonmastery groups

(3) norms and passing rates

(4) empirical methods for discovering standards

(5) empirical methods for adjusting cutoff scores, given a standard on an external criterion measure.

Most of the methods are based on decision theory (Hambleton & Novick, 1973; Subkoviak, 1980); the object is to maximize decision validity once the external criterion is established. The dichotomy of the test is matched to the dichotomy of the criterion while minimizing the number of false-positives and false-negatives. Huynh (1976) also presents a reasonable procedure for setting the cut-off score once the external criterion is established. Millman (1973), Shepard (1980), and others have suggested techniques for adjusting the cut-off score to protect against the more serious type of error. In some applications (certification, promotion, etc.) a false-positive could be considered a much more serious error than a false-

negative. Some of the other methods which reduce errors on both sides of the cut-off score are discussed in more detail in the validity and reliability sections.

Shepard (1980) also suggests criteria for selecting the appropriate standard-setting method for specific uses. The types of information needed is organized into three categories based on the level of decision-making that will be done. The three categories are: (1) individual diagnosis, (2) individual certification, and (3) program evaluation. Each of the three types would require a different strategy for establishing the cut-off score.

Validity

"A test has validity if it measures what it purports to measure" (Allen & Yen, 1979, p. 95). Any test, whether it is CR or NR must be valid in order to allow the user to attatch meaningful interpretations to the measurements provided. Several methods can be used to assess validity, depending on the test and its intended use. The two major types of validity are content validity and construct validity. The recommended procedures for assessing each of the these types of validity for CR tests are similar to the traditional NR approaches; however, because correlation coefficients, factor analysis, and multitrait-multimethod analysis all require a large variance, new statistics and new interpretations of the old statistics have been developed (Berk, 1980b; Popham & Husek, 1969).

Content Validity. Content validity can be established through a logical examination of the content of a test in

relation to a domain or a set of objectives. It can be divided into two forms: face validity and logical validity, both which are subjective in nature. Face validity can be established by having an individual (anyone from an expert to an examinee) critically examine the test and conclude that it does measure the relevant trait. Logical or sampling validity, "involves the careful definition of the domain of behaviours to be measured by a test and the logical design of items to cover all the important areas of this domain" (Allen & Yen, 1979, p. 96). Content validity is quite often the only type of validity that is addressed by makers of CR tests (Popham & Husek, 1969; Linn, 1980). Because the development of CR technology has increased the precision with which the test items are linked to the domain or list of objectives, many tests are constructed without mention of the content validity. Linn (1980) and others have suggested that many test constructors, assume that the validity of a criterion-referenced test is guaranteed by the definition of the domain and the process used to generate the items. This is further confirmed by Popham and Millman (in Berk, 1980c) in their discussion of domain specifications, item generation forms, and amplified objectives. All three techniques have been developed to ensure that the items are truely reflective of the domain being tested.

In order to establish a certain level of content validity for CR test items, Hambleton et al (1978) have suggested that the general approach involves judgements of test items by content specialists. Berk (1980b) has summarized three methods for obtaining these subjective evaluations of the content

validity. The first method is the item-objective congruence as discussed in the earlier section on item analysis and selection. The second method is to use a rating scale, in which, each judge will rate the items subjectively on their content in relation to the domain being tested. The last method involves having the judges try to match each of the items to a specific objective from the domain being tested. All three of these methods can be used to demonstrate the content validity of a CR test.

Construct Validity. Construct validity is much more difficult to establish and is often not even reported for a CR test. A test's construct validity is "the degree to which it measures the theoretical construct or trait that it was designed to measure" (Allen & Yen, 1979). One of the possible reasons can again be associated with the low variance in the test scores from CR exams. The homogeneous results are partly due to the fact that the CR tests are generally administered to a group, just prior to, or just after, a series of lessons on the domain being tested. This low variance deters the test maker from using the traditional methods of establishing construct validity such as factor analysis and multitrait-multimethod analysis. Instead, once the content validity of the items is ensured, the test constructor should concentrate on establishing the construct validity through the use of an external criterion. This involves decision or criterion related validity and is established through empirical methods which include setting a cut-off score and testing predictions. As mentioned earlier, test validity is closely related to the cut-off score. A cut-off score that is too high or too low will result in too many

false-negatives or false-positives, respectively, which will put the construct validity in doubt.

Berk (1976) suggests that an alternative method for identifying the optimal cut-off score is to compute a validity coefficient for each possible cut-off score. The subjects would be dichotomized according to an external criterion (e.g., uninstructed=0, instructed=1) and then their test scores would be dichotomized (a score below the cut-off score=0, a score at or above the cut-off score=1). The validity coefficient is simply the Pearson correlation (phi coefficient) between these two dichotomous variables. A test with high validity would result in a high positive correlation. Therefore, the test constructor would select the optimal cut-off score at a point where the validity (correlation) coefficient would maximized. Allen & Yen (1979) refer this procedure as demonstrating concurrent validity which is very similar to predictive or criterion related validity.

Reliability

A CR test has reliability if it shows "consistency of decision making across parallel forms of the criterion-referenced test or across repeated measurements" (Hambleton & Novick, 1973, p. 166). Relative to validity, much more attention has been given to the reliability of CR test construction. More than a dozen indices have been proposed by various researchers, and critical reviews have been conducted by Berk (1980a), Hambleton, Swaminathan, Algina, & Coulson (1978), Millman (1980), and Subkoviak (1978). Although these indices

have been available for a number of years, they are seldom reported for CR tests. Usually only a Kuder-Richardson 20 or 21 is reported for the internal consistency of the test. The various indices that are available have been placed, by Hambleton et al. (1978), into categories based on the similarities in their assumptions and methodology. The three categories, as shown in Appendix E, are threshold loss, squared-error loss, and domain score estimation.

Threshold loss. Threshold loss is defined by Berk (1980b) as "the consistency of mastery-nonmastery classification decision-making across repeated measures with one test form or on parallel test forms." The concept of using a threshold loss function was first proposed by Hambleton and Novick (1973, p. 168). Its use assumes (Berk, 1980b):

a) a dichotomous, qualitative classification of students as masters and nonmasters of an objective based on a threshold or cutting score.

b) the losses associated with all false mastery and false nonmastery classification errors are equally serious regardless of their size.

Several authors have proposed variations on the original index proposed by Hambleton and Novick. In total, six approaches have been suggested and are listed in Appendix E. All of the suggested indices are proportions based on one of two basic indices, "po, proportion of individuals consistently classified as masters and nonmasters across (classically) parallel test forms, and K, proportion of individuals consistently classified beyond that by chance" (Berk, 1980a). There are obvious

advantages and disadvantages for using each of these indices depending on the context of the decisions being made with the test scores. Berk (1980a) discusses each of the six indices, in detail, in the context of the test being used for individual decisions, for certification, and for program evaluation.

Squared-error loss. Squared-error loss functions are based on the squared deviations of the individual scores from the cut-off scores. This is similar to the threshold loss functions except that it builds in a sensitivity to the degree of mastery or nonmastery. Misclassifying examinees who are far above or below the cut-off score is considered more serious that misclassifying those who are near the cut-off score. Also the degree of error associated with false-positives and false-negatives is not assumed to be equal. Only two indices are available in this category, one proposed by Livingston (1972,1977) and one by Brennan (1980). Both indices are very similar in their assumptions and even in their calculations. Again, see Berk (1980a) for a detailed discussion of their common features as well as their slight differences.

Domain score estimation. Domain score estimation involves the use of statistics which are concerned generally with estimating the stability of an individual's score or proportion correct in the item domain independent of any mastery standard (Berk, 1980a). Five statistics, as listed in Appendix I, have been proposed by Brennan (1980), Berk (1980c), Cochran (1963), Lord (1959), Lord & Novick (1968), and Millman (1974). All of these statistics can be used in the context of individual or programmatic decision making and are calculated from parallel or

randomly parallel test forms. The main utility for this approach is in situations where a cut-off score is not available or is not necessary for that particular application. A detailed discussion of the actual indices is beyond the scope of this paper and can be found in Berk (1980a).

Score Interpretations

The interpretation of the test scores is listed as the final stage of test construction in Table 1. In actual fact, many of the earlier procedures are dependent on the context in which the scores will be interpreted. Score interpretation is not only important when the test is complete, but also essential throughout the developmental stages. Some of the interpretations that can be made include; (1) referencing the test score to an objective, a domain, or a cut-off score in order to assign pass/fail grades for minimum competency in a certification process, (2) to monitor an individual's progress through an educational sequence or, (3) to make decisions regarding program effectiveness. Many decisions made during the construction of a CR test, such as selection of appropriate validity and reliability procedures, are based on the context of the test score interpretations. For example, a test score that will be used solely for the purpose of assigning pass/fail grades should not employ reliability or validity procedures which are dependent on several cut-off scores. The interpretation of the test scores should be established at the start of the test construction and used as a guide throughout the developmental stages.

Structure and Purpose of Fitness Instructor Certification Program

Organizational Structure of Sport Governing Bodies

The BFTE that was constructed as part of this study was a CR test for the British Columbia Recreation Association-Fitness Branch. Because the exam was developed specifically for the BCRA-Fitness Branch it was necessary to operate within the framework and philosophy of this organization. Thus, it is appropriate to describe the administrative arrangement and terms of reference for the sport governing bodies associated with the project.

The overall authority for initiating, evaluating, and monitoring any fitness oriented projects in Canada is Fitness Canada. A subcommittee for Fitness Canada is the National Committee on Fitness Leadership Training in Canada which is directly responsible for the training of fitness leaders in Canada. Another group closely affiliated with Fitness Canada is the Canadian Association of Sport Sciences (CASS). CASS has a subcommittee, the Committee on Fitness Appraisal and Accreditation (CFAA), which has developed the registration model to be implemented throughout Canada. Both of these subcommittees work jointly with each other as well as with the sport governing bodies within each province.

In British Columbia, the sport governing body is Recreation and Sport Branch, British Columbia, which gets its funding from, and is, therefore, accountable to the Government of British Columbia. Working under the authority of the Recreation and

Sport Branch, B.C. is the British Columbia Recreation Association (BCRA), and more specifically the British Columbia Recreation Association-Fitness Branch. Therefore, the BCRA-Fitness Branch is responsible for the implementation of the nationwide fitness programs in British Columbia.

The BCRA-Fitness Branch established the Provincial Fitness Advisory Committee (PFAC) to provide the leadership and guidance during the implementation of the above programs. The members of the PFAC were appointed by the BCRA-Fitness Branch and included representatives from the following groups or organizations:

Sports Medicine Council
British Columbia Government
British Columbia Medical Association
British Columbia YMCA-YWCA
Canadian Association of Sport Sciences
Private sector
Corporate community
British Columbia Recreation Association-Fitness Branch

An overview of this structure of sport governing bodies can be seen in Appendix F.

The School of Physical Education and Recreation of University of British Columbia (UBC) dealt jointly with the BCRA-Fitness Branch and the PFAC only during the test construction phase of the fitness instructor certification program implementation. UBC acted as consultant on test

construction procedures and methodology, providing instruction and guidance during all stages. UBC provided all statistical and computer analyses at the item, subtest, and total test levels. Originally the BCRA-Fitness Branch had a three month timeline proposed for test construction, but later adopted the UBC suggestion for a nine month timeline to ensure a reasonable quality exam. This timeline is shown in Table 2. It should be stressed that the end result was not a UBC test--the BCRA-Fitness Branch made all final decisions (usually after consulting with UBC).

Purpose of Certification Program

The sudden increase in the popularity of fitness classes resulted in a drastic shortage of qualified, experienced fitness instructors. Both federal and provincial sport governing bodies were becoming concerned about the high number of unqualified fitness instructors. Coincidently, the number of injuries, from minor muscle stretches to torn ligaments, and stress fractures, had also increased dramatically. Both federal and provincial sport governing bodies became concerned about the high number of unqualified fitness instructors, and assumed that many of these injuries would be preventable by taking the proper precautions and following sound training principles. It was also assumed that by better preparing the instructors the participants can also develop appropriate training habits and, therefore, minimize injuries.

It is not uncommon for a participant to become an instructor after several months of attending fitness classes.

Variable experiences serve as a basis for the knowledge that is learned and will be passed along to any classes that he/she teaches, regardless of the worth of the information. As a result, many myths, as well as truths, are perpetuated very quickly without anyone questioning their validity. For example, until recently, it was not uncommon to observe groups of people performing situps in pairs with one partner sitting on the other's straight legs for more leverage. It is well known, by informed professionals, that this procedure is very detrimental to the lower vertebrae because of the excessive force applied to this area as a result of the situp motion in this position. Bent-legged situps or partial curl-ups, without a partner, is the prescribed procedure with this exercise. Another example of a harmful exercise is included in many warm-up procedures; the whole group is led through a systematic rotation their heads in clockwise and counterclockwise circles in an attempt to stretch the mucles around the neck. In actuality, these rotations provide very high stress, through the grating of the small vertebrae found in the neck. The prescribed procedure for the neck muscles has been shown to be side-to-side and forward-and-back motions rather than rotations.

Therefore, the major purpose of the certification is to provide increased minimum content and more uniform body of knowledge, better instruction, and a reduction in the number of injuries in the uninformed consumer.

Overview of Certification Process

The fitness instructor certification program was to be

implemented through several phases. The first phase involved establishing accredited courses, for fitness leaders, with the same set of objectives throughout the province. The next phase involved having fitness leaders completing the accredited course and then applying to the PFAC for certification. The PFAC would review the applicant with respect to their experience, completion of the accredited course, and then would do an on-site evaluation of the teaching abilities. After this, the applicant would have to pass the Basic Fitness Theory Exam (BFTE) in order to receive certification. The BFTE is the criterion-referenced multiple choice test developed within this study. The last phase of the program was for the fitness center, itself, to be evaluated in order to establish accredited fitness centers throughout British Columbia. These would be judged on various criteria, one of which was qualifications of instuctors.

The end result would be accredited fitness centers with proper facilities and certified instructors.

CR Test Construction Within the BCRA-Fitness Branch Framework

As stated previously, all final decisions regarding the test were made by the BCRA-Fitness Branch with input from the PFAC and UBC. UBC acted primarily as a consultant and advisor throughout the test construction process. The steps outlined earlier (in the CR Testing Overview section) were used primarily as guidelines for this process.

The first restriction the BCRA-Fitness Branch placed on the test was to insist that the test would be less than one hour in

length. They also wanted a multiple choice test for ease of marking. With the high number of applicants expected to write the test, the BCRA-Fitness Branch wanted to keep the administrative and marking costs to a minimum. The last restriction involved the passing or failing of the examinees. The examinees would pass or fail the test based on the total test score only. The subtest scores would not be considered, and thus an examinee could completely fail one subtest provided that he/she did reasonably well on the other two subtests.

Test Construction Procedures and Results for the BFTE

Overview

Table 2 below shows the general timeline for the procedures which were proposed for the construction of the BFTE. It is based on the first column of Table 1 which listed the developmental stages for CR test construction as suggested by Berk (1980b). Table 2 also shows that the CR test construction is basically an iterative process which continues indefinitely (pilot #1, two, three, ... etc.); a "finished" test must be continually monitered, and when necessary, updated to the current application (population, test use, etc.). This section of the paper deals with the construction procedures that were actually followed and the results obtained while constructing the BFTE.

Table 2

Proposed Time Line for the Procedures

Date	Procedure
Sept. 15 to Dec. 15	Domain specification Priorization and grouping of specific areas Construct specifications for item generation and mail to committee members
Jan. 12	Collect items from committee members
Jan. 26	Construct pilot #1 (feedback from committee)
Jan. 31	Final revisions of pilot #1
Feb. 2	Administer pilot #1 (two UBC classes; N=200)
Feb. 12	Computer analyses (LERTAP statistical package) Item and factor analyses Item selection and revision
Feb. 20	Construct pilot #2 Distribute for feedback Select a cut-off score
Feb. 25	Final revisions of pilot #2
Feb. 28	Administer pilot #2
Mar. 25	Computer analyses Item and factor analyses Reliability Validity Item selection and revision
Mar. 31	Present results to the PFAC and the BCRA-Fitness Branch
Apr. 30	Complete "final" draft (pilot #3)

Pilot Test #1

Identification of Learning Outcomes

Once the BCRA needs had been established and quality control issues resolved the actual test construction began. The first stage involved defining the domain of knowledge (i.e., the objectives) to be tested. In Table 1 this stage is referred to as content domain specification.

As the test was to be administered to students who had just completed the forty-hour instructor's course, this stage of the project involved defining the content of this course. The course content was specified in terms of main learning objectives which were categorized into eight "specific areas." These were contained in the "Fitness Instructor Criteria Prospectus" which was provided by the PFAC on behalf of the BCRA. The "specific areas" to be tested included:

1. Planning
2. Basics of Anatomy and Physiology
3. Safety
4. Excercise Principles for Adult Fitness Classes
5. Learning Theories
6. Teaching Strategies
7. Leadership
8. Evaluation

Ranking Specific Area. Because of the large number of "specific areas", with each containing many "main objectives" the members of the PFAC were asked to rate the relative importance of the eight "specific areas." The members were asked to subjectively rate the three most important and the the three least important "specific areas." A copy of the rating sheet that was used can be found in Appendix G. Eleven members responded and after tabulating the results a clear order was established. From the most to least important, the areas were ranked as in column one of Table 3.

Table 3

Priorization of Specific Areas

Rank	Area	New Category No.	Number Of Items
1.	Principles of Adult Fitness Classes	1	12
2.	Basics of Anatomy and Physiology [including Nutrition]	2	12
3.	Safety	3	9
4.	Teaching Strategies		
5.	Leadership	4	8
8.	Learning Theories		
6.	Planning		
7.	Evaluation	5	7
			Total=48

Because of the consistently low rating given to categories four through eight, and also the subjective comments provided by the judges, it was decided that these "specific areas" should be grouped together to form two larger, more general categories. Teaching Strategies, Leadership, and Learning Theories were combined to establish the new fourth category, while Planning and Evaluation were combined for the new fifth category. The five new categories are shown column three of Table 3.

Table 3 also shows the proposed number of test items for each of the five new categories. This again is based on the relative importance of the five categories and also on the preference for a test length of approximately 45 items. It has been shown that one can assume that examinees require approximately one minute per item for multiple choice items with four options if the items are testing at or below the application stage in Bloom's taxonomy (Berk, 1980b). It has also been shown that the minimum number of items required for any specific area is approximately six (Berk, 1980b; Wilcox, 1981). These factors were all considered when the number of items for each subtest were determined.

Submitting Items. At this stage in the test construction, members of the PFAC were asked to submit items relating directly to "main objectives" within their chosen "specific areas." They were asked to pick "specific areas" about which they felt most knowledgeable. Each member was familiar with the list of "main objectives" in the Fitness Instructor Criteria Prospectus and was provided with guidelines for submitting items, as shown in Appendix A (Berk,1980b), which included five sample items. The

submitted items were to be categorized by "specific area" and "main objective" being tested. It was also important to stress that the correct answer be supplied (circle, asterisk, etc.).

Items were collected over several months with the anticipation that pilot number one would be constructed with approximately ninety-six items, twice the number needed for the final version. It is generally assumed (Berk, 1980b) that approximately half of the items used in the first pilot will have to be removed or drastically revised. If less than twice the needed number of items is used in a first pilot then the test constructor runs the risk of not having enough items in the final version. If more than twice the needed number of items are in the first pilot then the test constructor could run into administrational problems such as:

1. fatigue in examinees
2. motivation of examinees
3. time constraints (classes, subjects, etc.)
4. cost of adminstration (personnel invigilating, location, photocopying, data entry/analysis)

Over two hundred items were collected from nine members of the PFAC. The breakdown of the items received as they related to the five new categories was as follows:

Specific Category	Number of Items Submitted
1. Principles of Adult Fitness Classes	51
2. Basics of Anatomy and Physiology (+ Nutrition)	83
3. Safety	44
4. Teaching Strategies/Leadership/Learning Theories	13
5. Planning/Evaluation	17
	Total=208

A relatively large number of items had been submitted for categories one, two, and three, however, only a few questions had been submitted for categories four and five and most of these were either intuitively obvious (no valid distractors were possible) or inappropriate for the basic fitness leader. Following lengthy meetings with course instructors and PFAC members it was concluded that the questions in categories four and five did not reflect the contents of the forty-hour course for basic fitness leaders, and it was agreed that although the material in these areas is within the overall list of objectives, in actual fact, very little is taught in these areas during the limited time available in the course. Closer examination of the detailed table of contents showed that the basic fitness leader is not required to handle administrative procedures such as advertising, overall program planning and evaluation, and individual fitness appraisal and exercise prescription. The basic fitness leader is more concerned with evaluations, based on exercise and safety principles, done while

conducting a class. Also, the material in category four is much too complex to teach in a detail during a forty-hour course. Because of the subjective nature of the material, only a minimal treatment is possible in this time. Therefore, it was decided that the examinee's knowledge of materal in categories four and five would be assessed through the Individual Competency Evaluation (ICE), rather than the BFTE. The ICE would be an assessment of the examinees knowledge and skills in a practical situation. This would allow for a much more objective evaluation of the examinee's knowledge with the BFTE.

Based on the above decision, a new breakdown was derived for the proposed number of items in each of the three specific areas which were retained. The breakdown would reflect the relative importance of each category. As mentioned earlier, ideally, pilot #1 would consist of approximately ninety items. However, as the subject pool of examinees for pilot #1 were to be Physical Eduation students at The University of British Columbia, a maximum time limit of 50 minutes was imposed (this is the length of the classes). This is also implied that a maximum of 50 items should be used, based on the theory that approximately one minute is required for each item (Berk, 1980b). Having assumed that these students could continue to write the exam into the 10 minute break they have between classes, a 57-item test was constructed. It consisted of 20 items from category one (Principles of Adult Fitness Classes), 22 items from category two (Basics of Anatomy and Physiology, including Nutrition), and 15 items from category three (Safety).

Randomization Procedures

The items from each of the three categories were scrambled using a table of random numbers. This was done to ensure that there was no pattern in the selection of the questions from each category. After this was complete, the table of random numbers was again used to ensure that no specific pattern emerged for the correct response to each question. While many randomization procedures exist, a simple table of random numbers was employed for both randomizing the categories and the correct responses.

Validity

Having organized the randomization of the items and the correct responses, a 57-item first draft of pilot #1 was constructed. At this point, several PFAC members subjectively evaluated the test to provide information concerning its face validity. Face validity is "established when a person examines the test and concludes that it measures the relevant trait. The person making this examination can be anyone from an expert to an examinee" (Allen and Yen, p.96). A committee was then formed to make the final revisions to pilot #1. These changes were made based on the following considerations:

1. the information available concerning face validity
2. deviations from item construction guidelines
3. the need to correct and to ensure the consistency of the grammer throughout each item and the entire test

Pilot #1 of the test was now ready to be administered.

Administration

Two classes in the Physical Education program at the UBC were designated as the examinees. The first course, Physical Education 163 (Biodynamics of Physical Activity), is compulsory for students entering the Bachelor of Physical Education program. These students (Year I group) were assumed to be relatively uninstructed in the body of knowledge being tested through pilot #1. The pilot test was administered to this group within the first few weeks of enrolling in this course. The second course, Physical Education 391 (Human Functional Anatomy and Applied Physiology), is also compulsory, but is taken by students in their second year. These students (Year II) had completed one term of their course when they wrote the BFTE pilot test These two "diverse" groups of subjects permitted for an external variable to be used in evaluating the discrimination ability of each item and the test as a whole.

Several other factors were also considered to be important in the selection of these two classes. The factors included:

1. sample size -- both classes are relatively large (approximately one hundred students each)
2. background of the subjects --interest in physical fitness
3. convenience -- easy to arrange
   -- easy to administer
   -- location (able to provide consistent

adminstrative conditions,
instructions,etc.)

4. cost -- no delivery/return costs

5. time -- no delay in getting completed exams back

In total, pilot #1 was administered to 92 Year I and 72 Year II subjects.

Results

Subjective feedback. To acquire useful subjective feedback, direct questions were asked of the examinees (the questions which were included with pilot #1 are shown in Appendix D). This feedback was valuable in detecting ambiguous questions, difficult or inappropriate terminology, poor distractors, or many problems which do not become apparent through the analysis of the item responses. The differences and similarities between the subjective responses by each of the criterion groups served to highlight imperfections in item construction identified potential problem areas. The feedback, in conjunction with the statistical analysis, was used to make the changes to pilot #1. For example, 38 examinees (16 Year I and 22 Year II) indicated that they did not understand the word "varus" in question number forty-five. The data analysis showed that only 41% of all subjects answered this item correctly and that it discriminates poorly between the Year I and Year II subjects. As a result of the above, and following closer examination of the course objectives, as well as meetings with course leaders, the PFAC decided to replace the item.

It should be noted that many of the examinees commented

that it would have been more useful if the subjective feedback questionnaire was placed at the start of the exam instead of the end. This would allow the examinees to more critical as they could read each item and assess it according to the guidelines in the questionnaire. When it was presented at the end, many of the examinees could not answer the questions about the items very well because they did not remember where the problems had been encountered as they did the test.

Statistical analyses-Overview. The data collected from pilot #1 were analyzed on the UBC mainframe computer using the Laboratory of Education Research Test Analysis Package (LERTAP). The LERTAP package provides statistical information at a variety of levels, including total test, subtest, item, and individually for all four possible responses. It also provides subtest and total test scores for each of the examinees.

The output was used to do an analysis of pilot #1 at the item, subtest, and total test levels. As the total test was to be used to determine whether or not individuals have sufficient knowledge to be ceritified as basic fitness leaders, its most important characteristic is its ability to discriminate between those who are knowledgeable and those who are not. The overall test score must discriminate while minimizing the number of false-negatives and false-positives. The item level analysis is described in the following section (see Appendix H for definitions and more information on the interpretation of item statistics).

Statistical analyses-Item level. The item analysis results are described through the use of two examples of major types of

item statistics that can emerge; an item with a high p-value and an item with a low p-value for the correct response (Due to the limitations of space in this paper, only two types of items are considered). For each type, the discussion will consist of the following five parts:

1. label the type of item
2. present the original item
3. present the item statistics, including the subjective comments
4. identify the patterns and suggest revisions
5. present the new version of the item

High P-value (Q1, subtest EP.item #1)

		Correlations			Means		
Option	P-value	ST	TT	EC	ST	TT	EC
* 1	96.3 *	0.04	0.10	0.04	* 14.43	34.14	0.44 *
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	2.4	-0.03	-0.06	0.02	14.00	31.75	0.50
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Other	1.2	-0.02	-0.09	-0.10	14.00	29.50	0.00

1. Cardiorespiratory endurance can be best defined as:
   * a) The efficiency of the heart and lungs
   b) The ability to sprint 100 metres in 10 seconds
   c) Vital capacity plus residual volume
   d) The mobility of the joint

No subjective comments were made by the Year I group. One Year II examinee referred to the item as too easy.

The high p-value (96.3) suggests that perhaps the item was too easy for both the Year I and Year II subjects. This is reconfirmed by the low (almost zero) correlation with the external criterion and by the mean value for the external criterion of 0.44. Both of these statistics suggest that the correct answer was chosen by approximately equal numbers of Year I and Year II subjects. The fact that virtually everyone selected the correct response also explains the low correlation between the correct answer and both the subtest scores and the total test score. A much higher degree of variance in the scores is necessary before a reasonably high correlation is

possible. Berk (1980b) and Allen and Yen (1979) have suggested that an ideal p-value is between 40.0 and 70.0.

Distractors two and four both have p-values of zero, indicating that nobody selected them as possible correct answers. As a result, both of them were replaced. Distractor three was selected by 2.4 percent ($\underline{n}$ =6) of the subjects. Unfortunately, four of these were from the Year II group. The negative, but small correlations with subtest and total test scores indicate that the subjects choosing this distractor are less knowledgeable. Therefore it was assumed that the four Year II subjects who chose distractor three were actually less knowledgeable in this area. This is also shown by comparing the total the total test mean (31.75) for the subjects who chose distractor three with those who chose the correct answer 9 (34.14). This indicates that distractor three is helping to discriminate. That is, these four Year II subjects are not really being classified as false-negatives.

After consideration of the subjective comments and the items statistics, distractors two (b) and four (d) were replaced. The new version of the item was ready.

1. Cardiorespiratory endurance can be best defined as:
   * a) The efficiency of the heart and lungs
   b) Stroke volume times heart rate
   c) Vital capacity plus residual volume
   d) Cardiac output minus residual volume

Low P-value (Q8, subtest AP.item #4)

		Correlations			Means		
Option	P-value	ST	TT	EC	ST	TT	EC
1	36.6	-0.13	-0.06	-0.11	12.72	33.55	0.37
* 2	36.0 *	0.32	0.26	0.28 *	14.66	36.07	0.63 *
3	17.1	-0.23	-0.21	-0.14	11.68	31.29	0.29
4	8.5	-0.02	-0.03	-0.05	13.07	33.43	0.36
Other	1.8						

8. Ulnar deviation means:

a) To turn the wrist outwards

* b) To turn the wrist inwards

c) To turn the wrist up

d) To turn the wrist down

One subject in the Year II group indicated that the item was confusing. No other comments were made by subjects from this group. Two subjects from the Year I group also found the item confusing. Two others felt there was either no answer or more than one answer. Eight subjects from the Year I group suggested that the terms "ulnar deviation" were unfamiliar and presented difficulty.

The correct answer had positive correlations with the subtest, total test, and the external criterion. It also had subtest and total test means which were much higher than for each of the distractors. All of the distractors had negative correlations with the subtest, total test, and the external criterion. The correlations for distractor 4(d) were not

significantly different from zero. That is, it is the only option that does not appear to discriminate among the knowledgeable and those with little or no knowledge in this area. However, even distractor 4(d), as do the other distractors, attracts more Year I that Year II subjects. This was shown by the mean external criterion value of 0.36.

In general, it could be stated that this item discriminates quite well among not only the knowledgeable and those without sufficient knowledge, but also among the Year I and Year II subjects. However, after the PFAC (revision committee) was presented with the above findings they recommended that the entire item should be replaced. They felt that even though the item discriminated well, the low p-value (36.0) suggested that it was too difficult. That is, they felt that too many false-negative decisions were made. It was also felt that the item was testing for knowledge of terminology that was beyond the scope of the forty-hour course. Knowledge of the term "ulnar deviation" was not stressed as an important objective for the course. The item was replaced by the following item:

8. In anatomical position, the bone located medially in forearm is called the:

   a) radius

   * b) ulna

   c) tibia

   d) fibula

The item replacement here re-emphasizes the fact that the item statistics did not entirely dictate the item revision process; but were just used to guide the test constructor in the decision-making process.

Table 4 shows a comparison of the p-values between the Year I and Year II groups for each of the items in pilot #1. For items which discriminate well, the p-value is considerably higher for the Year II group than for the Year I group. For example, item number #19 discriminates well, showing a difference of 60.9 between the Year II (75.0) and the Year I (14.1) groups. An example of a poor discriminator is item #25 which has p-values (Year I=34.8 and Year II=34.7) which are virtually equal. The items have been labelled, for easy identification of the potential problems. Those labelled with a "*" indicate that the Year II group does have a higher p-value than the Year I group, but by only five or less points. Those labelled with a "?" indicate an item on which the Year I group actually has a higher p-value than the Year II group. Both of these types of situations with p-values can warn the test maker of items which may be ambiguous, incorrectly keyed, poorly worded,

Table 4

Item p-values for Pilot #1

Subtest	Item	Year I	Year II		Subtest	Item	Year I	Year II	
EP	1	95.7	97.2	*	S	30	54.3	73.6	
AP	2	85.9	94.4		EP	31	60.9	61.1	*
AP	3	62.0	43.1	?	EP	32	76.1	84.7	
S	4	10.9	19.4		EP	33	65.2	66.7	*
S	5	12.0	11.1	?	S	34	6.5	15.3	
S	6	20.7	43.1		S	35	48.9	47.2	?
AP	7	60.9	84.7		AP	36	87.0	95.8	
AP	8	23.9	51.4		S	37	57.6	65.3	
AP	9	87.0	88.9	*	AP	38	8.7	66.7	
EP	10	91.3	97.2		EP	39	73.9	91.7	
S	11	60.9	68.1		AP	40	63.0	94.4	
S	12	40.2	40.3	*	EP	41	87.0	95.8	
EP	13	95.7	95.8	*	AP	42	35.9	31.9	?
AP	14	84.8	83.3	?	AP	43	38.0	22.2	?
EP	15	97.2	73.6	?	AP	44	25.0	66.7	
AP	16	34.8	100.0		AP	45	41.3	40.3	?
AP	17	59.6	80.6		EP	46	93.5	97.2	*
AP	18	32.6	54.2		S	47	27.2	48.6	
AP	19	14.1	75.0		AP	48	63.0	40.3	?
S	20	25.0	13.9	?	EP	49	75.0	90.3	
EP	21	47.8	69.4		EP	50	91.3	97.2	
S	22	17.4	25.0		S	51	66.3	58.3	?
EP	23	45.7	56.9		S	52	77.2	84.7	
AP	24	94.6	88.9	?	AP	53	54.3	68.1	

EP	25	34.8	34.7	?	EP	54	31.5	34.7 *
EP	26	51.1	73.6		EP	55	68.5	62.5 ?
EP	27	95.7	90.3	?	S	56	66.3	72.2
AP	28	87.0	77.8	?	EP	57	28.3	41.7
AP	29	55.4	44.4	?				

1. Year I (N = 92)
2. Year II (N = 72)
3. ?---Year I p-value is greater than or equal to Year II
4. *---Year II p-value is greater than Year I by 5 or less

or contain poor distractors. Table 4 also shows which subtest the items belong to with the symbols EP (Exercise Principles for Adult Classes), AP (Anatomy and Physiology plus Nutrition), S (Safety). This allows for a quick identification of a subtest which may contain many poor items. Overall, 8 items were labelled with a "*", and 17 items were labelled with a "?". This shows that, based on p-values, at least 25 out of 57 items needed further examination for possible revisions or deletions.

Statistical analyses-Subtest(ST) and Total Test(TT). The ST and TT statistics provided very useful information for the test construction process. These included means, standard deviations, low score, and high score for each subtest and for the total test as shown in Table 5 for pilot #1. The Year II group had a higher mean score than the Year I group on the total test (65% and 56%, respectively), and on each of the subtests. This can be interpreted as lending support to the validity of the test as the majority of the subjects would be categorized accurately by the total test scores. Subtest three (Safety) appeared to be the most difficult, or at least it had the lowest success rate, with both groups scoring under fifty percent. This indicates that some problems exist either in the items, in the list of objectives, or in the assumptions about these subjects. Closer examination of the item level statistics showed that there were problems with the Safety items.

Table 5

Descriptive Statistics for Pilot #1

	Year I	Year II	Combined
Total Test			
# of observations	92	72	164
mean	31.76=56%	36.92=65%	34.02=60%
s	4.98	5.64	5.85
low	20	26	20
high	45	48	48
# of items	57	57	57
Subtest #1: Exercise Principles			
mean	13.86=69%	15.13=76%	14.41=72%
s	1.91	2.37	2.21
low	11	9	9
high	18	19	19
# of items	20	20	20
Subtest #2: Anatomy and Physiology			
mean	11.99=55%	14.93=68%	13.28=60%
s	2.93	2.81	3.22
low	5	8	5
high	20	20	20
# of items	22	22	22

	Year I	Year II	Combined
Subtest #3: Safety			
mean	5.91=39%	6.86=46%	6.33=42%
s	1.88	2.18	2.07
low	2	3	2
high	10	13	13
# of items	15	15	15

The dispersion of scores, as represented by the standard deviation, was very similar for both groups in all three subtests and for the total test. Also, no major differences were detected between groups on the low or high scores in each category.

Table 6 also deals with subtest and total test statistics. It gives the correlations between subtests, total test, and an external criterion (Year I or Year II) for pilot #1. All correlations are positive, as expected, which shows that there are no major problems with the subtests. All correlations with the total test are 0.67 or higher for the combined groups, which again supports the reliability of the total test score as well as the subtests. The lowest correlations are with subtest three (Safety). This also supports the hypothesis, presented above, that some of the Safety items need revision or replacement. The correlations with the external criterion are also positive and significant thus lending support to the validity of the test. They are placing most people in the same categories as the groupings in the external criterion variable.

The above statistics for pilot #1 provides additional evidence for the validity and reliability of the test. If negative or zero correlations had been found then major revisions may have been necessary. This suggests that the care taken in the earlier stages of the test construction had proved to be worthwhile.

Table 6

Correlations for Pilot #1

Combined Groups

	EP	AP	S	TT	EC
EP	1.0	0.50	0.19	0.72	0.29
AP		1.0	0.44	0.90	0.46
S			1.0	0.67	0.23
TT				1.0	0.44
EC					1.0

Year I Group

	EC	AP	S	TT
EP	1.0	0.37	0.05	0.62
AP		1.0	0.42	0.89
S			1.0	0.64
TT				1.0

Year II Group

	EC	AP	S	TT
EP	1.0	0.52	0.21	0.76
AP		1.0	0.36	0.86
S			1.0	0.66
TT				1.0

Item revision or replacement was done with a variety of factors taken into consideration. Revision and replacement of the item stems, distractors, correct responses, and sometimes even the complete item, were systematically considered. The 57 items in pilot #1 varied in their patterns for p-values, correlations, and mean scores. The resulting large number of item statistics patterns make it impossible to discuss the decisions that were made on each one of the fifty-seven items. However, the examples provided (and Appendix H) should provide sufficient evidence for the types of decisions that were made.

It should be noted at this point that the statistical analyses were used, only as a guide to item revision or replacement, as item statistics should not be used as the sole basis for the decision-making process. Other factors needed to be considered before a decision, for change or against it, is made. Several of the factors are discussed below.

Test conditions can greatly influence the statistics that will result. For example, if one group has plenty of time to do the test and another group is rushed, the results can be very different. Pilot #1 was administered under similar condition to both groups; the subjects had no previous warning of the test and were given 50 minutes (in a classroom) to complete the test. Along with the test conditions, the appropriateness of the subjects can also influence the way the results need to be interpreted. In pilot #1 the Year II group turned out to be only slightly more knowledgeable than the Year I group. This also reflected the attitude of the students in the Year II group; these second year students had already been subjected to

numerous research projects. Also, the Year II group appeared to be only slightly more knowledgeable because of the specificity of the terminology in the area being tested.

The purpose of the test and the eventual interpretation of the test scores are also factors that can influence decisions regarding item revision or replacement. In the case of this test, the results were to be used as a basis for certification of Fitness Instructors. As the certification program was to be implemented on a volunteer basis, any decision made was greatly influenced by the ability of the test items to discriminate with a very low number of false-positives and false-negatives. Too many false-positives would have resulted in certified instructors without adequate knowledge; this implies that the objectives of the program to improve the quality of instruction and to reduce injuries would not be met. On the other hand, too many false-negatives would have resulted in many frustrated subjects who did have sufficient knowledge, but were unable to become certified. This would result in a very quick collapse of this volunteer certification program.

Another factor that influenced the decision-making process during test revisions was the knowledge of the experts in the area being tested. In some cases there was a subjective desire, by one or more experts, to include an item without revisions, even though the statistical analysis showed that problems existed. As noted earlier, UBC acted as consultants in the test construction process; the final decisions were made by the BCRA-Fitness Branch. Several items from pilot #1 which resulted in "poor" statistics were retained as a result of expert judgement.

Some of these items were later discarded or revised after pilot #2, while some of them showed "better" statistics with pilot #2.

All of the above factors were accounted for during test revisions. Also, to eliminate individual biases, the test revisions were done by a committee which consisted of several individuals who were experts in the subject matter (PFAC members) and one member which had enough statistical knowledge to interpret the computer printouts and guide the committee.

This concluded the analysis of pilot test #1. In all, 22 items were revised, 22 items were replaced, no items were deleted, 3 items were added, and 13 items were left unchanged. The first draft of pilot test #2, a 60-item test was constructed.

Pilot Test #2

Item-objective Congruence

The next stage in the validation procedure consisted of eight members of the PFAC were asked to judge how well each of the items could be matched to a particular objective from the forty-hour course. Each judge was asked to rate the item-objective congruence for each of the 60 items. The domain being tested was specified by giving the name of the subtest to which the item belonged. Space was also provided beside each rating for any additional comments. The Item-objective Congruence sheet which each reviewer received can be found in Appendix C. The reviewers were also asked to complete the Subjective Feedback questionnaires that the examinees had completed at the end of pilot #1. The results of the ratings, in conjuction with the subjective comments, both provided by the judges, were used to construct the final draft of pilot #2. The test consisted of 60 items.

Administration

The sample used for administering pilot #2 was more representative of the specific target population than the subjects used for pilot #1; these subjects were all intending to become certified fitness instructors. The test was administered at various locations throughout British Columbia at which the accredited forty-hour course for fitness instructors was being taught. Some groups wrote the exam at the end of the course, while others wrote the exam at the commencement of the course.

These groups were then put into two categories, Year I and Year II, respectively. This external criterion provided an even more diverse group than in pilot #1 and, therefore, can be considered a good test of the exam's discriminating ability. Overall, pilot #2 was administered to 200 subjects, 94 Year I and 106 Year II.

Statistical Analyses

Statistics were obtained through the LERTAP package, as for pilot #1, the items were examined at the item, subtest, and total test levels with respect to p-values, correlations, and means. A discussion of these indices is omitted at this point because of the similarity to those in pilot #1. The actual values of these indices can be found in Appendix I, Tables I-1, I-2, and I-3. However, as well as repeating the same statistics, several new analyses were also done to help determine the reliability of the test. These were postponed until this stage in the test construction because of the need to initially establish the validity the test.

Reliability

Many different reliability indices have been suggested by various authors (see Appendix E). Because of the similarities between many of these indices and the readily available access to LERTAP, only three indices were used here. These are: Hoyt's Estimate of Reliability, Cronbach's Composite Alpha, and the Standard Error of the Measurement (SEM). Each of these indices are defined and discussed below. Others, such as the Kappamax

and the P , are discussed in more detail earlier in this paper (also see Berk, 1980a).

The three indices were examined at this point to establish reliability within the subtests and for the total test. Another estimate of reliability which was also used here is the examination of the correlations between the subtests themselves and with the total test. For a discussion of this procedure refer back to pilot #1.

Hoyt's Estimate of Reliability. Hoyt's r is an estimate of the internal consistency of the total test, which ignores the subtests and treats the items as equal, but separate, entities. The values calculated for pilot #2 are all positive and significant as shown in Appendix I, Table I-1, with the estimate of reliability for the whole test being 0.86. This implies that in general the items are testing one body of knowledge. Estimates for the individual subtests are also positive and greater than 0.45, which is somewhat low, but not when it has resulted from a CR test with less variance than a traditional test. The lack of variance, especially in the subtests which have only 15 to 20 items, restricts the possible range of the reliability index.

Standard Error of Measurement. The SEM is an estimation of the standard deviation that would occur "for a specific examinee over repeated independent testings with the same test or parallel tests," (Allen and Yen, p.88). It refers to the observed test score as having a true part and an error part, with SEM being an estimate of the error portion in the observed

score. In pilot #2 the SEM is only 3.29 for the total test with both groups combined. This value is quite small considering there are 60 items in the test. It shows that 68% of the time, the interval formed by an individual's observed score plus/minus 3.29 will contain the individual's true score. The SEM for all three subtests are also relatively small compared to the number of items in each subtest. Again, this demonstrates the small errors in measurement; that is, it lends support to the relatively high level of reliability, especially for the total test.

Cronbach's Composite Alpha. Cronbach's Composite Alpha is only available for the total test, not for the subtests. It does not consider the item level differences as in Hoyt's index; instead, only the subtest scores are considered as separate entities and an index is calculated to show how well the subtests hold tegether. That is, the index shows to what degree the subtests are measuring a similar body of knowledge. The value of this index for pilot #2 is 0.78. As with the other measures, this also supports the reliability of the test.

Overall, these three indices show support for the total test score reliability. The subtests have a varying degree of reliability with the most reliable being being subtest two (Anatomy and Physiology) and the least reliable being subtest three (Safety). These results also correspond to the correlations between the subtests themselves and between the subtests and the total test as shown in Appendix I, Table I-2. Subtest three correlates lowest (0.77) with the total test and it correlates lower with each of the subtests (subtest one(EP) =

0.53 and subtest two(AP) = 0.62) than the correlation between the other two subtests (0.70). The combination of the correlations and the reliability indices demonstrate the reliability of the total test score. This concluded the analysis of pilot #2 and a committee was again formed to revise pilot #2 based on these results. In all, only one item was revised, no items were replaced or added, 15 items were deleted, and 44 items were left unchanged. Pilot #3, a 45 item test, was constructed.

Pilot Test #3

Pilot #3 consisted of only 45 items. Deletion of 15 items ensured that only the best discriminating items needed to be retained. This test conformed to the original restriction to keep the test to under fifty items.

Cut-off Score

The number 45 was not randomly chosen, but was carefully selected by the PFAC during the final phase of the test construction in which UBC was involved. This phase involved setting a cut-off score for the total test that would determine whether or not a subject passed or failed. Appendix J shows a table that helped the committee in their decision-making process. It shows, based on the sample size and statistics from pilot #2, the number of false-negatives and false-positives that would result with different cut-off scores, from 50% to 90%, for tests with 60, 46, and 36 items. This is similar to a procedure described earlier in this paper and also in Berk (1980a), where he suggests calculating validity coefficients to find the optimal cut-off score.

For all three test lengths the table shows that a low cut-off score, such as 50%, results in more false-positives than a higher cut-off score of approximately 80%. A low cut-off score means that most people will pass and that many may not actually have sufficient knowledge. The opposite is true if the cut-off score is set too high; very few will pass and as a result, many subjects who have the knowledge, will still not receive certification. There are also the previously mentioned practical aspects, such as ensuring enough people become

certified, rather that frustrated, so than this volunteer certification program will not collapse. After several meetings with the PFAC and the BCRA-Fitness Branch it was decided that a false-positive was more serious than a false-negative, because this would result in certification of the unqualified fitness instructor that the program is trying to eliminate. On the other hand, the majority of the committee felt that the public would not accept such an extremely high cut-off score as 75% or higher. Both the 60 and 36-item test would require this in order to keep the number of false-positives lower. The 46-item test would keep the false-positives low even at a 65% cut-off score (only two false-positives), however, the PFAC felt that the public would infer that the test was too easy. Therefore, the cut-off score was set at 70% which results in the same number of false-positives, but many more false-negatives than 65% (82 to 69). The test length was set at 45 items (one item was deleted just to round off the number) for pilot #3. For the 45-item test a passing score of 70% indicates 32 correct responses.

Results

Pilot #3 was constructed, in conjunction with the PFAC, after a statistical and subjective analysis of the data obtained from the administration of pilot #2. The statistics derived were similar to those performed on pilot #1, with the addition of the reliability indices and the setting of the cut-off score. In all, 15 items were deleted from pilot #2, one item was revised, and no items were added. The various indices showed strong support for the reliability of the total test score and

the cut-off score was set at a point which demonstrated maximum validity. It was concluded that pilot #3, a 45-item CR test, with a cut-off score of 32 (70%), was ready to be used in the Basic Fitness Leader certification program.

Summary and Recommendations

Summary

Pilot test #3 was the "finished" CR test that the BCRA-Fitness BRanch started using in the certification process for fitness instructors. This ended UBC's involvement in the project and the BCRA-Fitness Branch was satisfied that pilot #3 was sufficiently valid and reliable. A copy of pilot #3 can be found in Appendix K.

The procedures followed were close to the ones proposed in Table 2, with the exception of the time line being extended slightly due to unforeseen delays. For easy reference, a summary of the actual test construction procedures that were implemented in this project can be found in Table 7. A summary of the changes in the items (revisions, replacements, etc.) is shown in Table 8. For example, Table 8 shows that item number one was revised after pilot #1, but then was not changed for pilot #3. Some items (e.g., 3,4, and 11) were revised or replaced when pilot #2 was constructed and then deleted for pilot #3. Other items, such as 14,16, and 28, were written well enough for pilot #1 and did not have to be changed for pilot #2 or pilot #3. Overall, when pilot #2 was constructed from the analysis of the data obtained from pilot #1, 22 items were revised, 22 items were replaced, no items were deleted, 3 items were added, and only 13 items were left unchanged. When pilot #3 was constructed from the analysis of the data obtained from pilot #2, only one item was revised, no items were replaced or added, 15 items were deleted, and 44 items were left unchanged.

The figures demonstrate the general improvement in the quality of the test items. The number of test items from pilot #1 to pilot #2 was increased from 57 to 60, respectively, and then reduced to 45 for pilot #3.

Table 7

Summary of Actual Procedures

List of objectives recieved from the BCRA-Fitness Branch

Rating and ranking of Specific Areas (reduced from 8 to 3)

Guidelines for submitting items mailed out

- -items generation forms

Items submitted

Constructed first draft of Pilot #1

- -randomized items and correct responses
- -presented to the PFAC for feedback

Constructed Pilot #1 with PFAC

- -correct grammer
- -item revision

Administered Pilot #1 at UBC (Physical Education)

- -92 subjects in Year I
- -72 subjects in Year II
- -subjective questionnaire attatched to test

Statistical Analyses for Pilot #1

- -informal examinee feedback
- -item difficulty (p-values)
- -item discrimination
  - max possible gain (response means, r)
  - combined groups (item r with ST and TT)
  - item criterion (partial r)
- -descriptive statistics (ST and TT level)
  - mean
  - standard deviation
  - low, high scores

number of items

-correlations (ST, TT, and EC)

Constructed first draft of Pilot #2

Presented results to the PFAC

-subjective feedback

-item-objective congruence (content validity)

Constructed Pilot #2 with PFAC

Administered Pilot #2

-94 uninstructed subjects

-106 instructed subjects

Statistical analyses for Pilot #2

-as for Pilot #1 plus the following

-set a cut-off score for TT mastery/nonmastery

-reliability

Hoyt's r

S. E. M.

Threshold loss errors (false-neg/pos)

-validity

criterion-related

decision validity

Presented results to the PFAC

Constructed Pilot #3 with the PFAC

Presented Pilot #3 to the BCRA-Fitness Branch

Table 8

Summary of Item Revision from Pilot One to Three

Pilot Number			Pilot Number		
One	Two	Three	One	Two	Three
1	REV	NC	31	REV	NC
2	REV	NC	32	REV	NC
3	REV	DEL	33	REV	NC
4	REP	DEL	34	REV	NC
5	REP	NC	35	REP	DEL
6	REV	NC	36	NC	NC
7	REV	NC	37	NC	NC
8	REP	NC	38	REP	NC
9	REV	NC	39	REV	NC
10	REP	NC	40	REV	DEL
11	REV	DEL	41	REV	NC
12	REP	DEL	42	REP	NC
13	REV	NC	43	REP	NC
14	NC	NC	44	REP	NC
15	REP	NC	45	REP	NC
16	NC	NC	46	REP	DEL
17	REV	NC	47	REP	NC
18	REP	NC	48	NC	DEL
19	REV	NC	49	REP	NC
20	REP	NC	50	REV	NC
21	NC	DEL	51	REV	NC
22	REP	NC	52	NC	NC

23	REP	NC	53	NC	NC
24	NC	DEL	54	REP	DEL
25	REV	REV	55	REV	NC
26	NC	DEL	56	REP	DEL
27	REV	NC	57	NC	NC
28	NC	NC	58	ADD	NC
29	NC	NC	59	ADD	NC
30	REP	DEL	60	ADD	DEL

# of items in last pilot	57	60
# of items in new pilot	60	45
# of items revised (REV)	22	1
# of items replaced(REP)	22	0
# of items deleted (DEL)	0	15
# of items added (ADD)	3	0
# of items not changed (NC)	13	44

Future plans for the BFTE include substituting a few similar, but new, items into the test periodically with a constant monitoring of the test score validity and reliability. In this way, it is hoped that a large pool of valid and reliable items will be generated. The plan also includes a desire to increase the item pool until it is large enough to generate equivalent forms of the test randomly. To date, only a few new items have been added to the item pool and the BFTE has been administered to about one hundred subjects, with ninety of these people recieving certification as a Basic Fitness Leader Level I. The future of the BFTE and the implementation program look very optimistic.

Recommendations

Although the test has been given to the BCRA-Fitness Branch for use in the certification model this does not imply that the test construction process is complete. The test needs to be continually monitored to ensure that it is valid, reliable, and that the cut-off score is still appropriate. A change in the number of false-positives or false-negatives may indicate that the cut-off score needs to be adjusted.

Future projects of this nature need to ensure several points in order to maintain control of the test construction procedures and thereby result in a quality product. Firstly, the list of objectives should be developed in conjunction with the test constructor in order to establish a well-defined domain of objectives that can be expressed in behavioural terms and measured at the appropriate cognitive level (domain

specifications). This allows the test constructor to establish content validity during the early stages of the test construction. Secondly, the test constructor should maintain final approval over decisions regarding item revision and the setting of cut-off scores. If this is not the case, many other factors, such as the public's rejection of low cut-off scores in this project, can interfere with the decision-making process. And lastly, when an external criterion is to be used (such as instructed-uninstructed groups) to establish construct validity, the test constructor should ensure that two diverse groups actually exist.

References

Aiken, L. R. (1979). Relationships between the item difficulty and discrimination indexes. Educational and Psychological Measurement, 39(4), 821-824.

Allen, M. J. & Yen, W. M. (1979). Introduction to measurement theory. Monterey, California: Brooks/Cole Publishing.

Berk, R. A. (1976). Determination of optional cutting scores in criterion-referenced measurement. Journal of Experimental Education, 45, 4-9.

Berk, R. A. (1978). Consumer's guide to CRT item statistics. Measurement in Education, 9(1).

Berk, R. A. (1980a). A framework for methodological advances in criterion-referenced testing. Applied Psychological Measurement, 4, 563-573.

Berk, R. A. (1980b). A consumer's guide to criterion-referenced test reliability. Journal of Educational Measurement, 10, 159-170.

Berk, R. A. (1980c). Criterion-referenced measurement: The state of the art. Baltimore: John Hopkins University Press.

Block, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. Educational Researcher, 11(3), 4-11,plus 16.

Brennan, R. L. (1980). Applications of generalizability theory. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: John Hopkins University Press.

British Columbia Ministry of Education. (1979). British Columbia Assessment of Physical Education. Victoria, B.C.: Queen's Printer.

British Columbia Ministry of Education. (1980). Secondary Physical Education Curriculum and Resource Guide. Victoria, B.C.: Queen's Printer.

Canadian Association of Sport Sciences. Fitness appraisal certification and accreditation program. Manuscript in progress.

Chester, W. H., Aiken, M. C., & Popham, W. J. (Eds.). (1974). Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation,3.

Cochran, W. G. (1963). Sampling techniques (2nd ed.) New York: Wiley.

Cohen, J. (1983). The cost of dichotomization. Applied Psychological Measurement, 7(3), 249-254.

Cranton, P. A. (1976). An introduction to criterion-referenced measurement. Canadian Journal of Education, 1(4), 83-92.

Cronbach, L. J. (1971). Validation of educational measures. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington:American Council on Education.

Darst, P. W. & Steeves, D. (1980). A competency based approach to secondary student teaching in physical education. Research Quarterly for Exercise and Sport, 51(2), 274-285.

De Gruijter, D. N. M. & Hambleton, R. K. (1983). Using item response models in criterion-referenced test item selection. In R. K. Hambleton (Ed), Applications of item response theory (pp. 123-141). Vancouver, British Columbia: Hemlock Printers.

Eignor, D. R. & Cook, L. L. (1981). [Review of A criterion-referenced measurement model with corrections for guessing and carelessness]. Applied Psychological Measurement, 5(1), 137-139.

Fitzpatrick, A. R. (1983). The meaning of content validity. Applied Psychological Measurement, 7(1), 3-13.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. American Psychologist, 18, 519-521.

Glaser, R. & Bond, L. (Eds.). (1981). Testing: Concepts, policy, practice, and research. American Psychologist, 36(10).

Green, K. E. (1983). Subjective judgement of multiple choice item characteristics. Educational and Psychological Measurement, 43(2), 563-570.

Haladyna, T. M. & Roid, G. H. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. Journal of Educational Measurement, 18(1), 39-54.

Haladyna, T. M. & Roid, G. H. (1983). A comparison of two approaches to criterion-referenced test construction. Journal of Educational Measurement, 20(3), 271-282.

Hambleton, R. K. (1980). Contributions to criterion-referenced testing technology: An introduction. Applied Psychological Measurement, 4, 421-424.

Hambleton, R. K. (1983). Applications of item response models to criterion-referenced assessment. Applied Psychological Measurement, 7(1), 33-44.

Hambleton, R. K. (Ed.). (1980). Contributions to criterion-referenced testing technology. Applied Psychological Measurement, 4, 421-575.

Hambleton, R. K. (Ed.). (1982). Item response theory. Applied Psychological Measurement, 6(4), 373-492.

Hambleton, R. K. (Ed.). (1983). Applications of item response theory. Vancouver, British Columbia: Hemlock Printers.

Hambleton, R. K. & De Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 20(4), 355-368.

Hambleton, R. K. & Eignor, D. R. (1978). Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 15(4), 321-327.

Hambleton, R. K., Mills, C. N., & Simon, R. (1983). Determining the length for criterion-referenced tests. Journal of Educational Measurement, 20(1), 27-38.

Hambleton, R. K., Swaminathan, H., & Algina, J. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational

Research, 48, 1-47.

Hambleton, R. K. & Van der Linden, W. J. (Eds.). (1982). Advances in item responses and applications. Applied Psychological Measurement, 6(4), 373-473.

Huyhn, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Huyhn, H. (1977). The kappamax reliability index for decisions in domain-referenced testing. Paper presented at the annual meeting of the American Educational Research Association, New York.

Huyhn, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics, 4(3), 231-246.

Kane, M. T. (1982). A sampling model for validity. Applied Pschological Measurement, 6(2), 125-160.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. Applied Psychological Measurement, 4, 547-561.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 9, 13-26.

Livingston, S. A. (1977). Psychometric techniques for criterion-referenced testing and assessment, In J. D. Cone and R. P. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel.

Livingston, S. A. (1980). Comments on criterion-referenced testing. Applied Psychological Measurement, 4, 575-581.

Livingston, S. A. & Wingersky, M. (1979). Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement, 16(4), 247-260.

Lord, F. M. (1959). Tests of the same length have the same standard error of measurement. Educational and Psychological Measurement, 17,510-521.

Macready, G. B. & Dayton, C. M. (1980). The nature and use of mastery models. Applied Psychological Measurement, 4, 493-516.

Marshall, J. L. (1976). The mean split-half coefficient of agreement and its relation to other test indices: A study based on simulated data (Technical Report No. 350). Madison, WI: Wisconson Research and Development Centre for Cognitve Learning.

Martuza, V. R. (1977). Applying norm-referenced and criterion-referenced measurement in education. Boston: Allyn and Bacon.

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkley, CA: McCutchan.

Millman, J. (1979). Reliability and validity of criterion-referenced test scores. In R. E. Traub (Ed.), New Directions for testing and measurement: Methodological Developments. San Francisco, CA: Josey Bass.

Mills, C. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20(3), 283-292.

Nitko, A. J. (1980). Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 50(3), 461-485.

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response and scoring methods for multiple choice items: An empirical study of probabilistic and ordinal response models. Applied Psychological Measurement, 2, 83-96.

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Rogosa, D. R. & Willett, J. B. (1983). Demonstrating the reliability of the diffference score in the measurement of change. Journal of Educational Measuremant, 355-344.

Safrit, M. J. & Stamm, C. L. (1980). Reliability estimates for criterion-referenced measures in the psychomotor domain. Research Quarterly for Exercise and Sport, 51(2), 359-368.

Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Shiflett, B. & Schuman, B. J. (1982). A criterion-referenced test for archery. Research Quarterly for Exercise and Sport, 53, 330-335.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple choice items. Journal of Educational Measurement, 19(3), 211-220.

Stacking, M. L. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7(2), 201-210.

Stamm, C. L. & Moore, J. E. (1980). Applications of generalizability theory in estimating the reliability of a motor performance test. Reserch Quarterly for Exercise and Sport, 51(2), 382-388.

Stricker, L. J. (1982). Identify items that perform differentially in population subgroups: A partial correlation. Applied Psychological Measurement, 6(3), 261-274.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 13, 265-276.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 15(2), 111-116.

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: John Hopkins University Press.

Truab, R. E. & Rowley, G. L. (1980). Reliability of test scores and decisions. Applied Psychological Measurement, 4, 517-545.

Van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. Applied Psychological Measurement, 4, 469-492.

Van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-refenced test items. Review

of Educational Research, 51(3), 379-402.

Weltman, A. & Regan, J. (1982). A reliable method for the measurement of constant load maximal endurance performance on the bicycle ergometer. Research Quarterly for Exercise and Sport, 53(2), 180-183.

Wilcox, R. A. (1980). Determining the length of a criterion-referenced test. Applied Psychological Measurement, 4, 425-446.

Wilcox, R. A. (1981). [Review of Criterion-referenced measurement: The state of the art]. Applied Psychological Measurement, 5(1), 133-135.

Wilcox, R. R. (1979). On false-positive and false-negative decisions with a mastery test. Journal of Educational Statistics, 4(1), 59-73.

Wilson, G. (1980). The Construction of a Criterion-Referenced Physical Education Knowledge Test. Master's Thesis, University of British Columbia.

Yalow, E. S. & Popham, W. J. (1983). Content validity at the crossroads. Educational Researcher, 12(8),10-14, plus 21.

Appendix A

Guidelines for Submitting Items

1. The stem should be phrased as a direct question or an incomplete sentence.
2. All of the answers to a given item should be of similar length and grammatical form.
3. Distractors should be plausible enough to draw some responses.
4. Answers should not be of the form:

"all of the above"

"a,b, and d, but not c"

etc.

5. Words used for emphasis, or which negate a phrase, should be underlined.
6. Four responses should be provided for each item, only one of which is correct.
7. Each item should be representative of a specific "enabling objective."

Example of test items

1. The posterior iliac muscles are used mainly as:
   a) Rotators of the thigh
   b) Abductors of the thigh
   c) Adductors of the thigh
   d) Extensors of the thigh

2. The main purpose of aerobic exercise is to improve:
    a) Power
    b) Flexibility
    c) Endurance
    d) Coordination

3. Which of the following programs would use the Fartle training principle?
    a) Weightlifting program
    b) Gymnastics program
    c) Volleyball program
    d) Running program

4. Cardiac Output is the product of:
    a) Blood pressure x heart rate
    b) VO_2 max x O pressure
    c) Stroke volume x resting blood pressure
    d) Heart rate x stroke volume

5. Ligaments are used to connect:
    a) Bone to bone
    b) Cartilage to muscle
    c) Muscle to muscle
    d) Muscle to bone

Appendix B

Item Statistics

Informal examinee feedback

Item difficulty

Item discrimination (maximum between groups, minimum within groups)

- Maximum possible gain
- B index
- Internal sensitivity
- Combined groups (item-total r)
- Item criterion (partial r)
- Change-item r (correlation/multiple regression)

Item homogeneity

- Four levels of Chi square

Appendix C

Item-Objective Congruence

Reviewer: _____ Date: _____

First, read through the domain specifications and the test items. Next, please indicate how well you feel each item reflects the domain specifications it was written to measure. Judge a test item solely on the basis of the match between its content and the content defined by the domain specification. Please use the five-point rating scale shown below:

Poor	Fair	Good	Very Good	Excellent
1	2	3	4	5

Circle the number corresponding to your rating beside the test item number. The domain specifications have been abbreviated as follows:

(AP) Anatomy and Physiology
(EP) Exercise Principles for Adult Fitness Classes
(S) Safety

Domain	Test Item	Item Rating	Comments
EP	1	1 2 3 4 5	
AP	2	1 2 3 4 5	
AP	3	1 2 3 4 5	
S	4	1 2 3 4 5	
S	5	1 2 3 4 5	
S	6	1 2 3 4 5	
AP	7	1 2 3 4 5	
AP	8	1 2 3 4 5	
AP	9	1 2 3 4 5	
EP	10	1 2 3 4 5	
S	11	1 2 3 4 5	
S	12	1 2 3 4 5	
EP	13	1 2 3 4 5	
AP	14	1 2 3 4 5	

Domain	Test Item	Item Rating	Comments
EP	15	1 2 3 4 5	
AP	16	1 2 3 4 5	
AP	17	1 2 3 4 5	
AP	18	1 2 3 4 5	
AP	19	1 2 3 4 5	
S	20	1 2 3 4 5	
EP	21	1 2 3 4 5	
S	22	1 2 3 4 5	
EP	23	1 2 3 4 5	
AP	24	1 2 3 4 5	
EP	25	1 2 3 4 5	
EP	26	1 2 3 4 5	
EP	27	1 2 3 4 5	
AP	28	1 2 3 4 5	
AP	29	1 2 3 4 5	
S	30	1 2 3 4 5	
EP	31	1 2 3 4 5	
EP	32	1 2 3 4 5	
EP	33	1 2 3 4 5	
S	34	1 2 3 4 5	
S	35	1 2 3 4 5	
AP	36	1 2 3 4 5	
S	37	1 2 3 4 5	
AP	38	1 2 3 4 5	
EP	39	1 2 3 4 5	
AP	40	1 2 3 4 5	

Domain	Test Item	Item Rating	Comments
EP	41	1 2 3 4 5	
AP	42	1 2 3 4 5	
AP	43	1 2 3 4 5	
AP	44	1 2 3 4 5	
AP	45	1 2 3 4 5	
EP	46	1 2 3 4 5	
S	47	1 2 3 4 5	
AP	48	1 2 3 4 5	
EP	49	1 2 3 4 5	
EP	50	1 2 3 4 5	
S	51	1 2 3 4 5	
S	52	1 2 3 4 5	
AP	53	1 2 3 4 5	
EP	54	1 2 3 4 5	
EP	55	1 2 3 4 5	
S	56	1 2 3 4 5	
EP	57	1 2 3 4 5	
AP	58	1 2 3 4 5	
EP	59	1 2 3 4 5	
S	60	1 2 3 4 5	

In addition, complete 'subjective feedback' form found at end of test.

Appendix D

Subjective Feedback

Please respond to the following questions. Do not hesitate to provide additional comments about the entire exam or individual items.

1. Did any of the items seem confusing?

2. Did any of the items have no correct answer or more than one correct answer?

3. Did any of the words give you difficulty?

4. Did any of the items seem too easy?

5. How would you rate the difficulty of the entire exam from one (easy) to ten (hard)?

6. Provide any additional comments.

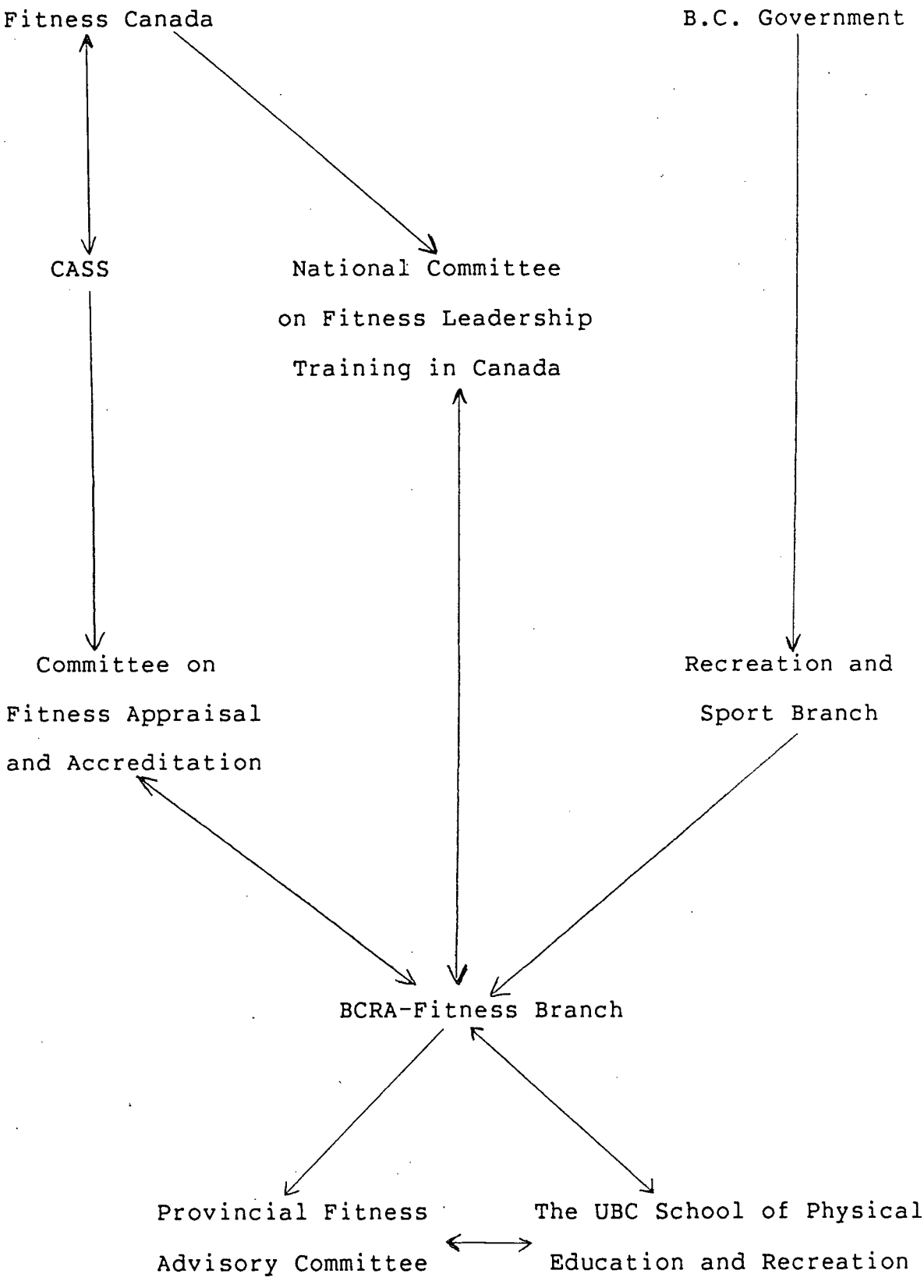
Appendix E

Reliability Procedures

Index	Source
Threshold loss	
P_o	(Hambleton & Novick, 1973)
K	(Swaminathan, Hambleton, & Algina, 1974)
β (P_o)	(Marshall, 1976)
P_o ,K	(Subkoviak, 1976, 1980)
P_o ,K	(Huynh, 1976)
KM	(Huynh, 1977)
Squared-error loss	
$K^2(X,Tx)$	(Livingston, 1972, 1977)
O(2)	(Brennan, 1980)
Domain score estimation	
Σp	(Berk, 1980; Cochran, 1963; Millman, 1974)
S.E. meas. (Xa)	(Lord, 1959)
	(Lord & Novick, 1968)
Φ	(Brennan, 1980)
	(Brennan, 1980)

Appendix F

Organizational Structure of the Sport Governing Bodies



Appendix G

Rating of "Specific Areas"

3 most important		3 least important
_____	A. PLANNING	_____
_____	B. BASICS OF ANATOMY AND PHYSIOLOGY	_____
_____	C. SAFETY	_____
_____	D. EXERCISE PRINCIPLES FOR ADULT FITNESS CLASSES	_____
_____	E. LEARNING THEORIES	_____
_____	F. TEACHING STRATEGIES	_____
_____	G. LEADERSHIP	_____
_____	H. EVALUATION	_____

Appendix H

Guidelines for Item Statistics Interpretations

Terminology

p-value --- For a given response, this is the percentage of examinees choosing this response. It indicates the item difficulty level. A higher p-value for the correct response implies that many subjects answered correctly; the item may be too easy. It should be noted that this paper will refer to the p-value of distractors as well as the correct answer. This is different from some authors who insist that the p-value is only applicable to the correct response. The p-value of distractors can provide invaluable information regarding the discrimination ability of the item.

item-subtest correlation --- For a given option, within an item, an examinee who chooses this option is coded as a one and all others are coded as zeros. Then a point-biserial correlation is calculated between these zero/one scores for this option and the examinees' scores on the subtest that contain this item. Because one variable is dichotomous and the other is continuous/multistep it is appropriate to calculate a point-biserial correlation. This procedure is repeated for all four of the possible options.

item-total test correlation --- The subjects are again coded as zeros and ones, for each option, and then these values are correlated with the examinees' total test scores.

item-external criterion correlation --- The subjects are coded as zeros and ones as above. They are then also coded on an external variable based on whether or not they were considered knowledgeable in this area. Examinees were coded as ones if they belonged to the instructed group and zeros if they belonged to the uninstructed group. A correlation was then calculated between the external criterion variable zero/one scores and the zero/one scores for each of four possible options.

item-subtest means --- For a given option, within an item, all examinees who choose this option will be included in the calculations. These examinees' scores, on the subtest to which the item belongs, are then averaged to find the mean score.

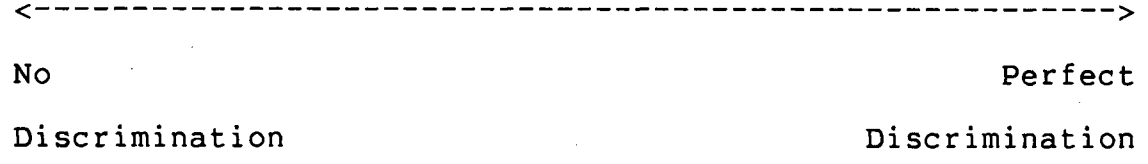
item-total test means --- For a given option, within an item, all examinees who choose this option will be included in the calculations. These examinees' total test scores are then averaged to find the mean score.

item-external criterion means --- For a given option, within an item, all examinees who choose this option will be included in the calculations. These examinees have been previously coded as one (instructed) or zero (unistructed) based on which group group they belong to. Then all of these zeros and ones will be averaged to fine the mean score. Guidelines

Item statistics obtained from the LERTAP package provide, for the test constructor, a good summary of the subjects' responses to each item. These statistics, combined with the

subjective feedback, are used to guide the decision-making process regarding item revision or replacement. Keeping in mind that the items will probably never be perfect in their discrimination ability in all populations, it is possible to place the items on a continuum based on their discrimination ability as shown below.

Discrimination Continuum



In reality, exams would seldom contain items at either end of the continuum; these are mainly theoretical reference points. In fact, most tests used today would contain items that fall somewhere along the continuum. For criterion-referenced tests it is important to attempt to develop items that are closer to the right-hand side of scale; this would be an item that discriminates without any false-negatives or false-positives. Table H-1 shows an example of the LERTAP output that would result from such an item. The table shows that the output would have the following characteristics:

1. The p-value (difficulty index) for the correct response

should be equivalent to the percentage of the total number of subjects which are instructed. This assumes that only the instructed will select the correct response. In Table H-1, 50% of the subjects were from the instructed group and the p-value is 50.

2. The p-value for each distractor should be equivalent to the percentage of the total number of subjects which are uninstructed divided by the number of distractors. In a four option test, the p-value should be about 16.7 (50/3) if equal numbers of instructed and uninstructed subjects were used. In Table H-1, it should be noted that the correlations are shown as 1.0 and -1.0 . In actual fact these values would be impossible under the conditions described. Because of the homogeneity of each of the groups, there could be no variance on each variable within groups. This implies that the correlation would be zero. If we, however, assume that the subtest and total test scores are also dichotomous (which they really are because all subjects score perfectly or get zero) then these extreme correlations are possible.

3. The correct response will have a high positive correlation with the scores from the subtest of which the item is a part. Table H-1 shows this to be a theoretical value of 1.0.

4. The distractors will each have a high negative

correlation (-1.0) with the scores from the subtest of which the item is a part.

5. The correct response will have a high positive correlation (1.0) with the total test scores.

6. The distractors will each have a high negative correlation (-1.0) with the total test scores.

7. The correct response will have a high positive correlation (1.0) with the external criterion.

8. The distractors will each have a high negative correlation (-1.0)with the external criterion.

9. The mean score on the subtest for those who chose the correct response should be equal to the number of items in the subtest. Table H-1 shows that the subtest mean to be a perfect score of twenty-two. This assumes that only the instructed subjects will choose the correct response and that all of the instructed subjects will choose the correct response.

10. For each distractor the mean score on the subtest for those who chose a given distractor should equal zero. Table H-1 shows subtest means of zero for all three distractors. The assumptions are similar to those in point nine.

11. For the correct response, the mean score on the total

test for those who chose this response should be equal to the number of items on the total test (57). Also see point number nine.

12. For each distractor, the mean score on the total test, for those who chose this distractor, should be equal to zero. Also see point number nine.

13. For those subjects who chose the correct response, the mean score on the external criterion should be equal to the code number assigned to the instructed group of subjects. Table H-1 shows that examinees from the instructed group were coded as ones.

14. For those who chose a given distractor, the mean score on the external criterion should be equal to the code number assigned to the uninstructed group. Table H-1 shows that examinees from the uninstructed group were coded as zeros.

NOTE: The above comments are made under the assumption that all other items in the test are also perfect discriminators.

Table H-1

A Perfect Discriminator

			Correlations			Means		
Option	N	P	ST	TT	EC	ST	TT	EC
* 1	90	50 *	1.00	1.00	1.00 *	22	57	1.0 *
2	30	16.7	-1.00	-1.00	-1.00	0	0	0.0
3	30	16.7	-1.0	-1.0	-1.0	0	0	0.0
4	30	16.7	-1.0	-1.0	-1.0	0	0	0.0
Total	180							

Table H-2

A Non-discriminator

			Correlations			Means		
Option	N	P	ST	TT	EC	ST	TT	EC
* 1	45	25 *	0.0	0.0	0.0 *	11	28	0.5 *
2	45	25	0.0	0.0	0.0	11	28	0.5
3	45	25	0.0	0.0	0.0	11	28	0.5
4	45	25	0.0	0.0	0.0	11	28	0.5
Total	180							

NOTE: 1. The items statistics are based on a sample size of 180. Of this total, 90 were considered to be instructed and 90 were considered to be uninstructed.

2. The subtest consisted of 22 items.

3. The total test consisted of 57 items.

No perfect discriminators were found during the item analysis of pilot one. While some items could be rated as good discriminators, many would fall closer to the other end of the continuum. A few could even be rated as being very close to the non-discriminating item. Table H-2 demonstrates some of the potential characteristics of the item statistics that would be derived from the non-discriminating item. These characteristics are listed below.

1. All possible responses should have an equal numer of subjects choosing it; that is, all of the options should have equal p-values. Table H-2 shows that each option received 45 of the total 180 (25%) possible responses. No distinction is made between the distractors and the correct response.

2. The correlations between the correct response and the subtest would be very low. Table H-2 shows this value to be zero. This implies that there is no relationship between whether the subject chooses the correct response and how they score on the subtest. A subject who chooses the correct response will not necessarily do well (or even poorly) on the subtest from which the item was taken. No accurate predictions can be made.

3. The correlations between each of the distractors and the subtest would also be very low. Again, Table H-2 shows the value to be zero. This implies that there is no relationship between which distractor a subject chooses and how well they score on the subtest.

4. The correlation between the correct response and the total test score would be very low.

5. The correlation between each of the distractors and the total test would also be very low.

Both 4 and 5 show that there is no relationship between the option that a subject chooses and the subject's total test score.

6. The correlation between the correct response and the external criterion would be very low.

7. The correlation between each of the distractors and the external criterion should be very low.

Both 6 and 7 are shown as zeros in table H-2. This demonstrates that there is no relationship between the option that a subject chooses and whether they were from the instructed group or not. This implies that instructed, as well as uninstructed, subjects were choosing all four options. No distinction can be made beween the distractors and the correct response based on the criterion groups.

8. The mean score on the subtest for those subjects who chose the correct response should be equal to about half the number of items in the subtest. Table H-2 shows the value to be eleven.

9. For each distractor, the mean score on the subtest, for those who chose the distractor is also equal to about half the number of items in the subtest.

For 9 and 10, Table H-2 shows the mean value to be eleven for a twenty-two item subtest. No discriminatin is made between the distractors and the correct response. The mean score being one-half of the total number of items reflects the fact that an equal number of instructed and unistructed subjects were selecting each option.

10. For the correct response, the mean score in the total test, for those who chose this option, should be equal to about half the number of items in the total test.

11. For each distractor, the mean score on the total test, for those who chose this distractor, should be equal to about half the number of items in the total test.

Note: For all of the above, assume that all of the other items in the subtest in total test are perfect (good) discriminators.

12. Assume that the instructed subjects were coded as ones and the uninstructed subjects were coded as zeros; for those subjects who chose the correct response, the mean score on the exterion criterion should be 0.5.

13. Assume that the instructed subjects were coded as ones and the uninstructed subjects were coded as zeros; for those subjects who chose a given distractor, the mean score on the external criterion should be equal to 0.5.

Points 12 and 13 both demonstrate the fact that half the subjects choosing any option (distractor or correct response) are instructed and the other half are unistructed.

APPENDIX I

STATISTICS FOR PILOT #2

TABLE I-1

Descriptive Statistics for Pilot #2

		Uninstructed Group	Instructed Group	Combined
Total Test	Number of Observations	106	94	200
	x	28.5 = 48%	38.2 = 64%	33.1 = 55%
	s	7.9	7.2	9.0
	Low	12	19	12
	High	48	52	52
	# of items	60	60	60
	Hoyt's r	0.82	0.81	0.86
	S.E.M.	3.34	3.61	3.29
	Cronbach's alpha	0.73	0.76	0.78
Subtest 1 Exercise Principles	x	11.8 = 56%	14.0 = 67%	12.9 = 61%
	s	2.9	2.7	3.0
	Low	4	8	4
	High	18	20	20
	# of items	21	21	21
	Hoyt's r	0.59	0.52	0.61
	S.E.M.	1.85	1.81	1.85
Subtest 2 Anatomy & Physiology	x	10.2 = 44%	15.7 = 68%	12.8 = 56%
	s	4.3	3.8	4.9
	Low	1	6	1
	High	21	23	23
	# of items	23	23	23
	Hoyt's r	0.75	0.72	0.82
	S.E.M.	2.13	1.99	2.07
Subtest 3 Safety	x	6.5 = 41%	8.4 = 53%	7.4 = 46%
	s	2.2	1.9	2.3
	Low	2	4	2
	High	13	13	13
	# of items	16	16	16
	Hoyt's Rel.	0.40	0.32	0.46
	S.E.M.	1.66	1.54	1.63

Table I-2
Correlations for Pilot #2

Combined Groups

	EP	AP	S	TT	EC
EP	1.0	0.70	0.53	0.85	0.37
AP		1.0	0.62	0.94	0.56
S			1.0	0.77	0.42
TT				1.0	0.54
EC					1.0

Uninstructed Group

	EP	AP	S	TT
EP	1.0	0.58	0.45	0.81
AP		1.0	0.54	0.91
S			1.0	0.74
TT				1.0

Instructed Group

	EP	AP	S	TT
EP	1.0	0.72	0.44	0.87
AP		1.0	0.47	0.93
S			1.0	0.68
TT				1.0

Table I-3

Item p-values for Pilot #2

SUBTEST	ITEM	UNIN[1]	IN[2]		SUBTEST	ITEM	UNIN	IN	
EP	1	94.3	90.4	?	EP	31	50.9	79.8	
AP	2	57.5	81.9		EP	32	51.9	80.9	
AP	3	34.9	48.9		EP	33	52.8	64.9	
S	4	24.5	39.4	*	S	34	8.5	8.5	?
S	5	50.0	79.8		S	35	59.4	52.1	?
S	6	41.5	46.8	*	AP	36	60.4	84.0	
AP	7	73.6	89.4		S	37	50.0	54.3	*
AP	8	49.1	87.2		AP	38	46.2	71.3	
AP	9	56.6	86.2		EP	39	76.4	90.4	
EP	10	90.6	94.7	*	AP	40	45.3	75.5	
S	11	59.4	69.1		EP	41	78.3	54.3	?
S	12	48.1	62.8		AP	42	34.0	42.6	
EP	13	90.6	95.7	*	AP	43	54.7	63.8	
AP	14	42.5	66.0		AP	44	56.6	86.2	
EP	15	72.6	89.4		AP	45	42.5	71.3	
AP	16	29.2	67.0		EP	46	10.4	44.7	
AP	17	43.4	58.5		S	47	46.2	63.8	
AP	18	20.8	57.4		AP	48	31.1	56.4	
AP	19	60.4	79.8		EP	49	20.8	48.9	
S	20	17.9	11.7	?	EP	50	94.3	92.6	?
EP	21	40.6	48.9		S	51	12.3	47.9	
S	22	42.5	93.6		S	52	77.4	98.9	
EP	23	44.3	68.1		AP	53	40.6	61.7	

SUBTEST	ITEM	UNIN[1]	IN[2]		SUBTEST	ITEM	UNIN	IN	
AP	24	45.3	75.5		EP	54	29.2	35.1	*
EP	25	28.3	33.0	*	EP	55	45.3	40.4	*
EP	26	50.9	76.6		S	56	11.3	75.5	
EP	27	88.7	93.4	*	EP	57	40.6	48.9	
AP	28	21.7	55.3		AP	58	45.3	53.2	
AP	29	28.3	55.3		EP	59	32.1	40.4	
S	30	84.9	92.6	*	S	60	15.1	11.7	?

1. UNIN - Uninstructed group (N=106)
2. IN - Instructed group (N=94)
3. ? - UNIN p-value is greater than or equal to IN
4. * - IN p-value is greater than UNIN by 5 or less

Appendix J

Number of False-Positives and False-Negatives

for Varying Cut-Off Scores

		60 Items		46 Items		36 Items	
		UNIN	IN	UNIN	IN	UNIN	IN
50%	F	60	14**	83	30	57	5
	P	46*	80	23	64	49	89
55%	F	75	24	91	38	64	10
	P	31	70	15	56	42	84
60%	F	87	35	99	60	72	16
	P	19	59	7	34	34	78
65%	F	94	47	104	69	85	23
	P	12	47	2	25	21	71
70%	F	97	60	104	82	92	30
	P	9	34	2	12	14	64
75%	F	103	72	105	93	93	37
	P	3	22	1	1	1	57
80%	F	105	86	106	94	104	66
	P	1	8	0	0	2	28
85%	F	106	90			105	79
	P	0	4			1	15
90%	F	106	94			106	94
	P	0	0			0	0

1. UNIN - Uninstructed group (N=106)
2. IN - Instructed group (N=94)
3. F - Fail
4. P - Pass
5. * - False positive
6. ** - False negative

1. Cardio respiratory endurance can be best defined as:

   a) The efficiency of the heart and lungs
   b) Stroke volume times heartrate
   c) Vital capacity plus residual volume
   d) Cardiac output minus residual volume

2. The most effective method for improving flexibility is through:

   a) Isotonic contractions
   b) Static stretches
   c) Isometric contractions
   d) Ballistic stretches

3. Heat exhaustion may be caused by all of the following except:

   a) Exercising while dehydrated
   b) Holding your breath while exercising
   c) Inadequate ventilation in room
   d) Exercising in multi-layered clothing

4. Stretching the hamstrings is important for low back problems because it:

   a) Strengthens the abdominals
   b) Re-aligns the spine
   c) Allows the pelvis to tilt backward
   d) Allows the pelvis to tilt forward

5. A significant proportion of lower back problems are directly related to a lack of muscular strength, particularly in the:

   a) Abdominal area
   b) Lower back area
   c) Upper back area
   d) Hamstring

6. The bones in your forearm are called the:

   a) Ulna and Tibia
   b) Radius and Ulna
   c) Fibula and Radius
   d) Tibia and Radius

7. The by-product of anaerobic work is:

   a) Glycogen
   b) Free fatty acids
   c) Carbohydrates
   d) Lactic acid

8. Which of the following is not considered a primary component of fitness:

   a) Muscular strength
   b) Flexibility
   c) Speed
   d) Cardio-vascular endurance

9. The best method for reducing body fat is to:

   a) Decrease dietary carbohydrate intake
   b) Increase your activity level
   c) Decrease dietary fat intake
   d) Decrease caloric intake and increase your activity level

10. Human skeletal muscle is composed of fast twitch and slow twitch fibers. Which one of the following is a characteristic of slow twitch muscle fibers?

   a) Functions by anaerobic metabolic processes
   b) Rich in glycolytic enzymes
   c) More resistant to fatigue
   d) Has a higher power output

11. While performing strength exercises, participants should be encouraged to:

   a) Breathe out during the relaxation phase
   b) Breathe out during the exertion phase
   c) Hold their breath to improve fitness
   d) Breathe only between repetitions of the exercise

12. The term that best describes the position of the lumbar vertebrae relative to the cervical vertebrae is:

   a) Superior
   b) Inferior
   c) Proximal
   d) Medial

13. When exercising at very intense levels the main fuel source is:

a) Fat
b) Protein
c) Glycogen
d) Free fatty acids

14. Your gracilis muscles are located:

a) In your upper back
b) In your lower leg
c) In your thigh
d) In your forearm

15. The group of muscles mainly responsible for hip abduction are the:

a) Hamstrings
b) Quadriceps
c) Gluteals
d) Abdominals

16. Sweating causes:

a) Temperatures in working muscles to increase
b) Salt concentration within you to increase
c) An increased blood flow to the working muscles
d) Oxygen utilization to be increased

17. Swelling can be limited BEST through:

    a) Cold treatments combined with elevation
    b) Moist heat treatments with aspirin
    c) Heat treatments combined with elevation and compression
    d) Cold treatments combined with elevation and compression

18. To lose a pound of body fat you'd have to experience a deficit of:

    a) 3200 calories
    b) 3500 calories
    c) 3800 calories
    d) 3000 calories

19. A person with high blood pressure in your fitness program should be told all of the following, except:

    a) Do not hold your breath on exertion
    b) Limit or delete strictly upper body exercises
    c) Avoid any extended isometric contractions
    d) Stand still and just do upper body and arm exercises during the cardio section of class

20. Doing a push-up on the hands and the feet instead of on the hands and the knees:

    a) Increases the workload
    b) Decreases muscular strength
    c) Provides better balance
    d) Decreases the resistance

21. Cardiac output is the product of:

a) Heart rate x blood pressure
b) Blood pressure x stroke volume
c) Stroke volume x heart rate
d) Heart rate x oxygen uptake

22. Which chamber of the heart receives the returning venous blood

a) Left artrium
b) Left ventricle
c) Right atrium
d) Right ventricle

23. Of the choices below, the best equation for approximating maximum heartrate is:

a) 220 - Age = Maximum Heartrate
b) 3 x Resting Heartrate = Maximum Heartrate
c) 220 - Resting Heart Rate = Maximum Heartrate
d) 170 - Age = Maximum Heartrate

24. In order to maximize gains in muscular strength one must increase the:

a) Duration of workouts
b) Frequency of workouts
c) Number of repetitions
d) Resistance of the workload

25. The most important concept in learning the proper mechanics of a sit-up is to:

a) Avoid using the arms to develop momentum
b) Curl up, with the back rounded
c) Keep the back straight
d) Anchor the legs for support

26. In treating a heat stroke victim, which of the following will best hasten the cooling process when applied in conjunction with sponging?

a) Applying ice cubes to the armpit and groin
b) Applying cold compresses to the neck and forehead
c) Giving the patient cold drinks
d) Fanning the patient

27. Among the following, all are sources of energy for exercise except:

a) Glycogen
b) Fats
c) Vitamins
d) Glucose

28. The first indication of oxygen shortage would be:

a) Cyanosis and dilated pupils
b) Difficulty in breathing and cyanosis
c) Dilated pupils and difficulty in breathing
d) Increased respiration and pulse rates

29. Glycogen is not stored in the:

    a) Stomach
    b) Muscles
    c) Kidney
    d) Liver

30. A warm-up for a moderate paced fitness class should not include

    a) Maximal isometric contractions
    b) Light jogging
    c) Light stretching
    d) Stationary walking

31. The main reason for doing joint mobilizing exercises as a warm-up activity is to:

    a) Prepare the musculoskeletal system for subsequent full range of motion activity
    b) Increase synovial fluid volume
    c) Increase joint cartilage thickness
    d) Dispose of waste products built up in the muscle cells

32. All of the following are considered to be connective tissue except:

    a) Tendons
    b) Ligaments
    c) Skin
    d) Collagen

33. Which of the following are the largest vertebrae in the spinal column?

a) Lumbar
b) Thoracic
c) Coccyx
d) Cervical

34. Lateral arm raises are associated with contraction of the:

a) Serratus Anterior
b) Deltoid group
c) Latissimus dorsi
d) Trapezius

35. The correct order of sections of the spine from top to bottom would be:

a) Thoracic, lumbar, cervical, sacral, coccyx
b) Cervical, thoracic, lumbar, sacral, coccyx
c) Coccyx, cervical, sacral, thoracic, lumbar
d) Cervical, sacral, thoracic, lumbar, coccyx

36. Pregnant women participating in your fitness class should:

a) Avoid doing pelvic tilt exercises
b) Avoid doing slow sit-ups
c) Avoid reaching their maximal heart rate
d) Avoid drinking fluids one hour prior to class

37. According to Canadian Food Guides, of your total daily nutritional consumption _____% should be comprised of carbohydrates:

a) 10-20%
b) 20-30%
c) 35-45%
d) 50-60%

38. Of the following progressions in locomotive cardiovascular training, which is the most appropriate for safe improvement at a beginner's fitness level?

a) Walking, to intermittent jogging, to continuous jogging to long slow distance running
b) Jogging, to interval speed work, to long distance running
c) Interval running, to varied pace running, to long slow distance running
d) Long slow distance, to interval fast running, to varied pace running

39. The initial assessment, a search for immediate life threatening problems, should be conducted in the following order:

a) Check the airway for obstruction
Check if they are breathing
Check if the heart is beating to circulate blood

b) Check the airway for obstruction
Check if the heart is beating to circulate blood
Check if they are breathing

c) Check if they are breathing
Check if the heart is beating to circulate blood
Check the airway for obstruction

d) Check if they are breathing
Check the airway for obstruction
Check if the heart is beating to circulate blood

40. Wearing plastic exercise apparel is:

a) Beneficial, because the increased perspiration causes greater fat loss
b) Beneficial, because it causes heat retention and thereby keeps the muscles warm
c) Not beneficial, because it causes a reduction in blood circulation
d) Not beneficial, because it does not allow the body heat to dissipate

41. The hamstring muscles are important for flexing the:

a) Hip
b) Knee
c) Pelvis
d) Ankle

42. A fifty year old man, considerably overweight, has been relatively sedentary for the past two decades. What should he concentrate on <u>first</u> in order to improve his fitness?

a) Muscular and cardiovascular endurance
b) Flexibility and muscular endurance
c) Flexibility and cardiovascular endurance
d) Cardiovascular endurance and muscular strength

43. When teaching "target heart rate monitoring", it is <u>most</u> important to emphasize that:

a) Maximal stroke volume increases with age
b) A lower resting heart rate implies a better fitness level
c) A heartrate of 170 beats per minute is better than one of 150
d) Maximal heart rate decreases with age

44. Which blood vessels carry nourishment to the heart muscle?

a) Coronary arteries
b) Coronary veins
c) Carotid arteries
d) Pulmonary artery

45. Residual lung volume does not increase with:

a) Age
b) Smoking
c) Exercise
d) Asthma