9-1982

Reliability of Selected Health and Performance Related Test Items from the Project Unique Physical Fitness Test Inventory

Gene A. Daquila
The College at Brockport

RELIABILITY OF SELECTED HEALTH AND PERFORMANCE RELATED TEST ITEMS FROM THE PROJECT UNIQUE PHYSICAL FITNESS TEST INVENTORY

A Thesis

Presented to the Professional Studies Graduate Unit
Division of Physical Education

State University of New York

College at Brockport

Brockport, New York

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Education

(Physical Education)

---

by

Gene A. Daquila

1982

STATE UNIVERSITY OF NEW YORK

COLLEGE AT BROCKPORT

BROCKPORT, NEW YORK

Professional Studies Graduate Unit
Division of Physical Education

Title of Thesis: Reliability of selected health and performance related test items from the Project UNIQUE physical fitness test inventory.

Author: Gene A. Daquila

Read and Approved by: Barry B. Shultz

Date submitted to the Professional Studies Graduate Unit: _____

Accepted by the Professional Studies Graduate Unit, Division of Physical Education, State University of New York, College at Brockport, in partial fulfillment of the requirements for the degree Master of Science in Education (Physical Education).

Date: Sept. 1982

Constance Koenig McGuigan
Chairperson of Professional Studies Graduate Unit

ACKNOWLEDGEMENTS

This thesis could not have been written without the help of numerous individuals to whom I wish to express my thanks.

First of all, I would like to thank my thesis committee. To my major advisor, Dr. Barry Shultz, I am extremely grateful for the time he shared with me, the knowledge he imparted to me, and the patience he extended me. To Dr. Frank Short, many thanks for his excellent advice, constant availability, and fine "relief work". My sincere appreciation goes to Dr. Joseph P. Winnick, for sharing with me his vast knowledge in the area of Special Physical Education, for his consistent support throughout the year, and for allowing me to use the Project UNIQUE data, which constituted a major part of this study.

Gratitude is expressed to those physical education teachers who helped in the collection of data and to the students who participated in this study. Also, a sincere thanks goes out to my fellow graduate assistants: Margaret Savage, Barbara Shields-Dildy, Lynn Rosenthal, Tracy Newlove, Joe Kelly, Kurt Clark, and Scott Connors, for their assistance in collecting data and their continued encouragement.

My deepest appreciation is extended to Angela Grossi for typing and proof reading this paper.

Finally, I am extremely grateful to my family and friends for their tireless support and constant encouragement.

COMPLETED RESEARCH IN HEALTH, PHYSICAL EDUCATION, AND RECREATION
State University College at Brockport (C. Koenig McIntyre)
Brockport, New York Institutional Representative

DAQUILA, Gene A. Reliability of selected health and performance related test items from the Project UNIQUE physical fitness test inventory.
M.S. in Ed. 1982; pp. 126.
(Dr. Barry Shultz)

This study was conducted to determine several psychometric qualities on selected items from the Project UNIQUE Physical Fitness Test Inventory. Coefficients of reliability (consistency within day and between day) were determined by intraclass techniques. The standard error of measurement was also determined for the selected items. In addition, the appropriate criterion score was determined by an analysis of variance and collaborated by an analysis of the superdiagonal of the inter-trial correlation matrix. For the multi-trial test items, 50 nonimpaired youth, 50 visually impaired youth, 50 auditory impaired youth, and 50 orthopedically impaired youth between the ages of 10 and 17, were randomly selected from the various schools participating in the Project UNIQUE study. Subjects for the single-trial items included 50 nonimpaired youth, 47 visually impaired youth, and 50 auditory impaired youth, between the ages of 10 and 17 randomly selected from at least two different schools in the Rochester, N.Y. and Buffalo, N.Y. areas. The results of this study showed that most of the test items were reliable. In addition, the reliability of the test items for the impaired groups was, in general, equal to or better than the reliability coefficients for the

nonimpaired group. The results also indicated that in most cases the Project UNIQUE scoring procedures were appropriate, although some changes were recommended.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The importance of physical fitness for the average American has grown tremendously in recent years. It has become quite common to see people jogging, bicycling, and engaging in other fitness activities in greater numbers than ever before. However, many physical educators have been emphasizing the importance of physical fitness for a healthy body and a healthy mind for some time (Winnick, 1979). Since Kraus and Hirschland (1954) published the results of their study, physical fitness has received a great deal of attention in the United States. In 1955, when President Eisenhower formed the President's Council on Physical Fitness, progress was made on improving the physical fitness of American children (Clarke & Clarke, 1978).

Within the last decade, the importance of physical fitness has received increased attention from special educators, psychologists, and physicians as they now support the need for physical development for the handicapped (Winnick, 1979). However, some students cannot be given physical fitness tests, either because their disabilities render them incapable of performing the tests or because their conditions may be aggravated in attempting them (Fait, 1978). Therefore, the impaired have often been neglected in terms of physical fitness programs. It is apparent that a need exists for physical fitness tests that can properly assess the fitness levels of the impaired.

Project UNIQUE has been developed as a possible aid in

assessing the physical fitness levels of impaired individuals. It is the intention of Project UNIQUE to determine the unique physical fitness and performance needs of sensory and orthopedically impaired individuals (Winnick & Short, 1980). Several health and performance related test items are being evaluated for their applicability to children and youth with handicapping conditions. From these several test items, an assessment instrument will be developed that will allow educators to assess and determine the health and fitness needs of sensory and orthopedically impaired populations (Winnick & Short, 1980).

In developing an assessment instrument such as the Project UNIQUE Test, one must be aware that there are certain characteristics or qualities which are essential to measurement. If these certain characteristics or qualities are lacking, little faith can be put in the measurement item, thus, little use can be made of it (Baumgartner & Jackson, 1975). One of the important qualities of a measurement tool is reliability. The reliability of a test means the degree to which one can expect the results to be consistent (Franks & Deutsch, 1973). Another way of thinking of reliability is whether a test measures the true average performance of an individual (Kirkendall, Gruber, & Johnson, 1980). Therefore, if an individual is measured twice with a perfectly reliable measuring device and if his/her ability has not changed, then the two scores should be the same. Taking it one step further, reliability refers to the dependability of scores. That is, if a test was administered on two occasions to the same students, the same differences between students would be detected (Safrit, 1981).

Statement of the Research Problem

The research problem for this study was to determine certain psychometric characteristics on selected health and performance related test items from the Project UNIQUE Physical Fitness Test Inventory. More specifically, an attempt was made to determine coefficients of reliability (consistency within day and between day) by using intraclass techniques. In addition, it was also intended to determine the standard error of measurement.

Sub-Problem

The sub-problem of this study was to determine the appropriate number of trials for selected health and performance related test items from the Project UNIQUE Physical Fitness Test Inventory. More specifically, the appropriate criterion score was determined by an analysis of variance and collaborated by an analysis of the suderdiagonal of the inter-trial correlation matrix.

Purpose of the Study

The purpose of this study was to determine several psychometric qualities on selected items from the Project UNIQUE Physical Fitness Test Inventory. Coefficients of reliability (consistency within day and between day) were determined by intraclass techniques. The standard error of measurement was also determined. In addition, it was intended to determine the appropriate number of trials and scoring method for the selected test items by considering trial-to-trial variation in the test scores.

It was expected that this study will be used to help justify the test items included in the final Project UNIQUE Test Battery

and the recommended number of trials. Furthermore, it was intended to contribute to the testing and assessment procedures of sensory and orthopedically impaired youth in the area of physical fitness.

Need and Significance

The need to develop quality programs and tests to determine the fitness levels of impaired individuals is clearly evident. The Project UNIQUE Physical Fitness Test Inventory is a step in this direction. This study will help establish the reliability of selected items from the UNIQUE test inventory. Thus, it should provide valuable information for physical educators assessing the physical fitness levels of impaired youth. It will aid researchers by presenting information in a needed area.

Franks and Deutsch (1973) point out that, "Reliability is important since the confidence a teacher has in being able to test the pupil's ability depends largely on the consistency of the test" (p. 12). In addition, reliability is important to both the test developer and the test user as it helps to identify the potential sources of variability among scores obtained for a group of individuals and to quantify the magnitude of this variability so as to improve the measuring device or testing schedule (Safrit, 1976).

In 1959 the Research Council of the American Alliance of Health, Physical Education and Recreation (AAHPER) indicated the importance of determining an appropriate number of trials for test items by stating: "The primary factor in obtaining consistent scores are adequate trials and objective scoring" (Scott, p. 112). Klesius (1966) reiterated this importance in

his study on reliability of the AAHPER Youth Fitness Test. This study will help establish the appropriate number of trials which should be used to achieve high reliability for the Project UNIQUE Physical Fitness Test Battery. This will increase the value of the test to Project UNIQUE staff and physical educators concerned with assessing the fitness level of impaired children.

Delimitations

1. For the multi-trial items, this study was conducted using 50 nonimpaired youth, 50 visually impaired youth, 50 auditory impaired youth and 50 orthopedically impaired youth from various schools throughout the United States that participated in the Project UNIQUE study.

2. For the single-trial items, this study was conducted using 50 nonimpaired youth, 47 visually impaired youth, and 50 auditory impaired youth from at least two different sites for each population. Sites were from the Buffalo, N.Y. and Rochester, N.Y. areas.

3. For the multi-trial items subjects were tested during the interval of April 1, 1980 and June 30, 1981.

4. For the single-trial items, subjects were tested during the interval of April 1, 1981 and June 30, 1981.

5. Ages of the subjects ranged from 10 to 17 years of age. Subjects were divided into two age categories, 10 to 13 and 14 to 17.

6. Project UNIQUE testing procedures and instructions were used for all test items as described in the Project UNIQUE Training Manual.

Limitations

1. Subjects tested on the single-trial items were randomly selected from predetermined sites. Therefore, this sample represented a smaller population than the subjects tested on the multi-trial items.

2. The sampling procedure did not take into consideration the sex of the subjects.

3. Visually impaired subjects were not further classified as to blind and partially sighted. Auditory impaired were not further classified as to deaf and hard of hearing. For orthopedically impaired, there was no further classification breakdown.

4. Various testers were used to collect data. All testers were certified by completing the Project UNIQUE training procedures.

5. No data was collected on orthopedically impaired students for the single-trial items due to the fact that a large majority of orthopedically impaired subjects do not perform the timed trunk taise and timed leg raise.

Definitions

Reliability. The tendency toward consistency exhibited by an individual's repeated performance of one behavior (Safrit, 1981).

Visually impaired. A visual condition which, after correction, adversely affects a child's educational performance. The term includes both partially sighted and blind children (Federal Register, December 30, 1976).

Auditory impaired. Includes the definitions of deaf and hard of hearing taken from the Federal Register of December 30, 1976. Deaf means a hearing impairment which is so severe that the child's hearing is nonfunctional for the purposes of

educational performance. Hard of hearing means a hearing impairment, whether permanent or fluctuating, which adversely affects a child's educational performance but which is not included under the definition of deaf.

Orthopedically impaired. For the purposes of this study the orthopedically impaired included four categories: amputees, congenital anomalies, spinal neuromuscular conditions and cerebral palsy, as defined by Winnick and Short (1980). Amputees are subjects who have part or all of one or more of their extremities missing. People with congenital anomalies include individuals whose extremities are fully or partially present and are deformed. Spinal neuromuscular conditions are primarily characterized by spinal lesions which directly affect limb functioning. Cerebral palsy is defined as a disorder characterized by disturbances in voluntary motor functioning resulting from lesions in the brain that affect the motor control centers.

Nonimpaired. Those children not identified as handicapped, who are free from physical impairments or disabilities which may influence test results and attend regular classes in non-institutionalized regular schools (Winnick & Short, 1980).

Multi-trial items. Those test items on the Project UNIQUE Test Inventory which were administered more than once. The multi-trial items being evaluated in this study are: skinfold measurements, rise to stand, mat creep, shuttle run, modified stork test, sit and reach, grip strength, flexed arm hang, standing broad jump and softball throw.

Single-trial items. Those test items on the Project UNIQUE

Test Inventory which were administered only once. The single-trial items being evaluated in this study are the timed leg raise and timed trunk raise. These items were administered only once due to the lengthy recovery time that would be required to insure maximum performance on a repeated trial.

CHAPTER II

REVIEW OF LITERATURE

In developing an assessment instrument such as the Project UNIQUE Test, one must be aware that there are certain characteristics or qualities which are essential to measurement. One of the more important qualities of a measurement tool is reliability. Reliability can be studied from two theoretical frameworks: classical test theory or generalizability theory. However, the practical nature of this study is not concerned with the theoretical derivations and issues which define these two approaches. Therefore, differences between the two theories will not be discussed.

In the sections that follow, a brief overview of reliability is presented. In addition, the different types of reliability are briefly discussed. Then the correlational procedures and analysis of variance procedures for determining a reliability coefficient are compared. Lastly, a review of related reliability studies conducted on the selected items is presented.

RELIABILITY

Two sets of measurements on the same individuals will never exactly duplicate each other. The fact that repeated sets of measurements never exactly duplicate one another is what is meant as unreliability (Stanley, 1971). At the same time, repeated measurements of a series of objects or individuals

will ordinarily show some consistency. This tendency toward consistency from one set of measurements to another is called reliability (Stanley, 1971). Safrit (1981) defines reliability as the tendency toward consistency exhibited by an individual's repeated performance of one behavior. Therefore, if an individual is measured twice with a perfectly reliable measuring device and if his/her ability has not changed, the two scores will be identical (Baumgartner & Jackson, 1975).

The reliability of a test refers to the dependability of scores or their relative freedom from error. According to Stanley (1971), "The evaluation of reliability of any measure reduces to a determination of how much of the variation in the set of scores is due to certain systematic differences among the individuals in the group and how much to other sources of variation that are considered, for particular purposes, errors of measurement" (p. 359). Safrit (1976) points out it is necessary to identify which portion of the total variance reflects the influence of systematic and/or constant factors and which portion reflects the influence of random or unpredictable factors.

Systematic or predictable factors can be thought of as those associated with "true" differences among individuals (i.e., sex, age, skill level), or with constant explainable "errors" arising from the individuals themselves (i.e., improvement over trials or sessions), or from the procedures and conditions that are a part of the process of administering the test and/or evaluating the test performance (i.e., a scale

that weighs each subject five pounds too heavy, or consistent differences between judges). Random error variance is caused by unexplained, unpredictable factors. It may be a composite of some factors within the individual, some factors in the measuring device, and/or some completely unknown factors. According to Safrit (1976), "It is the 'true' systematic differences between individuals that a test should be sufficiently reliable to detect" (p. 7).

Achievement of suitable reliability is dependent upon two basic factors: 1) reducing variation attributable to measurement error, and 2) detecting individual differences within the group measured, i.e. variation of the true scores. Thus, within the classical test theory approach, the reliability of an instrument is viewed in terms of its measurement error (error variance) and its power to discriminate different levels of ability within the group measured (true score variance). Safrit (1976) points out that reliability corresponds to the proportion of the total variance in test scores which is due to true differences among individuals in the quality being evaluated by the test.

Once determined, the coefficient of reliability can be evaluated in several ways: 1) by referring to standards in published tables, 2) by using the logical expectation for the type of skill being measured, and 3) by referring to the age and skill level of the students for whom the coefficient was determined (Safrit, 1981). Although there are no rigid standards for reliability coefficients, acceptable values are partially

determined by the situation in which the test will be used. Baumgartner & Jackson (1975) indicate that most physical fitness measures are quite stable from day to day, exhibiting test-retest reliability coefficients between .80 and .95. However, minimum acceptable reliability measurements should be based on the degree of reliability required and that which other individuals have obtained by testing similar students (Baumgartner & Jackson, 1975).

In their Standards for Educational and Psychological Tests (1974), the American Psychological Association indicates that reliability coefficients have limited value for test users and that the standard error of measurement (SE_m) is more useful. They report, "the standard error of measurement ordinarily is more useful than reliability coefficients; it has greater stability across populations...and it may be used to identify limits that have a defined probability of including the true score" (p. 50). Since the variance of all measurements contains some measurement error, the standard error of a test score reflects the degree one may expect a test score to vary due to measurement error (Baumgartner & Jackson, 1975). In addition, the standard error of measurement is presumed to be independent of the range of talent in the group for which it was determined giving a better indication of the absolute accuracy of measurement (Kroll, 1967).

Interpretation of the SE_m is especially useful when two or more students are compared on a test or when a student's score is being compared with a standard performance. The standard

error of measurement may be considered a standard deviation of a test score (Baumgartner & Jackson, 1975), thus, a student's score may be interpreted by utilizing the normal curve. Since we know the individual's obtained score and the $\underline{SE}_m$, we can be 68% certain that the true score will fall within one standard error of measurement above or below the obtained score (Safrit, 1981).

The standard error of measurement is of particular value when we are interested in applying the information with regard to consistency to different groups (Thorndike, 1951). Reliability coefficients depend upon the range of ability in the group from which the coefficients were determined. Thus, it is impossible to apply the coefficient directly to another group differing in variability on the trait in question or to compare results from different groups. The $\underline{SE}_m$, however, is usually independent of the exact spread of scores, therefore, it can be expected to remain uniform in groups of approximately the same level of ability (Thorndike, 1951). This means that it is possible to apply the value directly to new groups which may differ considerably in variability from the group on which the standard error of measurement was originally determined (Thorndike, 1951).

In general, the smaller the $\underline{SE}_m$ the more reliable the test. However, it must be pointed out that this is not always the case since the standard error of measurement is related to the magnitude of the standard deviation of the test. Safrit (1981) explains,

> A test with a standard deviation of 16 may have the same reliability as a test with a standard deviation of 8. However, the standard error of measurement of the first test will be numerically twice that of the second...Therefore, when comparing two different tests, the reliability coefficient should be evaluated with the magnitude of the standard error of measurement in mind. (p. 109).

Types of Reliability

Reliability can be classified into three types: stability, internal consistency, and equivalence. When individual scores change very little from one day to the next, they are said to be stable. If each person's score were to remain unchanged from one day to the next, then the scores would be perfectly stable and reliable. The test-retest method is used to obtain the stability reliability coefficient, which is an estimate of the measuring instrument's reliability (Baumgartner & Jackson, 1975).

With the test-retest method, each student is measured with the same test or instrument on two different days (day one and day two). The correlation between the day one and the day two scores is the stability reliability coefficient. The closer the coefficient is to unity (1.00), the more stable and reliable the scores are.

Many researchers use an internal consistency reliability coefficient as an estimate of the reliability of their measures (Baumgartner & Jackson, 1975). The advantage of this reliability coefficient is that all measures are collected in a single day. The term internal consistency refers to a consistent rate of student scoring throughout a test or, if multiple trials of a

test are administered, from trial to trial (Baumgartner & Jackson, 1975). According to the standards of the American Psychological Association (1974), "Estimates of internal consistency should be determined by matched-half or random-half methods or by analysis of variance procedures" (p. 53). It must be noted that matched-half coefficients reflect expert judgment and tend to be higher in value than random-half coefficients, while analysis of variance procedures tend to yield lower values than matched-half procedures (American Psychological Association, 1974).

The equivalence reliability coefficient, though seldom used with physical performance tests, is occasionally used with paper-and-pencil tests distributed on a national basis, i.e. the physical education subtest of the Graduate Record Exam (Baumgartner & Jackson, 1975). The equivalence reliability coefficient is obtained by the parallel forms method. Two tests of equal difficulty, which measure the same material (Form A and Form B), are developed. Each respondent then takes both Form A and Form B. The tests may be administered on the same day or on different days. By correlating the scores on the two forms the equivalence reliability coefficient is determined.

Although these terms still appear in measurement literature, Safrit (1981) indicates that the classification of coefficients into these three categories is no longer recommended. In their Standards for Educational and Psychological Tests the American Psychological Association (1974) reports that

> It is recommended that test authors describe the meanings of any coefficients they report as accurately and precisely as possible. It is informative to say, for example, "This

> coefficient indicates the stability of measurement of equivalent scores based on parallel forms of the test administered 7 days apart, without intervening practice or instruction." Although lengthy, such a description is reasonably free from ambiguity. (pp. 49-50)

Procedures for Estimating Reliability

There are many statistical techniques for estimating the reliability of tests. However, this study will limit its discussion to the correlation procedure and the analysis of variance (ANOVA) procedure. Test descriptions of most measures of motor skill developed prior to 1965 report reliability estimates determined by using the well-known Pearson product-moment correlation coefficient. Although this procedure has been widely applied to tests of motor performance as well as to written tests, other techniques are generally more appropriate for tests used in physical education (Safrit, 1981). However, because of the popularity of this approach it will be discussed before proceeding with the more appropriate ANOVA technique.

Pearson product-moment correlation. According to Safrit (1981), the Pearson product-moment correlation compares the relative positions of a group of individuals on two sets of scores. In order to compute a correlation coefficient for scores on two tests, the scores must be obtained for the same group of individuals on both tests. These scores are converted from raw scores to Z scores, so that direct comparisons can be made.

The correlation between two trials of the same test for the same individuals can also be obtained using the product-moment method. Since this computation is possible, the Pearson

product-moment correlation has often been incorrectly used to estimate the reliability of tests involving repeated measures. Safrit (1981) indicates that when two sets of scores are available for the same test, the conversion of raw scores to Z scores will mask any systematic increases or decreases from one set of scores to another. Therefore, using the interclass correlation coefficient to estimate the reliability of repeated measures may result in an inaccurate estimate.

There are several limitations in using the Pearson product-moment correlation coefficient to estimate reliability. The major inadequacy of the method, according to Kroll (1962) is that it does not differentiate from among several possible sources of error which may be involved in the measurement. Safrit (1976) points out that the product-moment correlation coefficient is a bivariate statistic and should be used to determine the relationship between two different variables, like height and weight. When subjects are tested twice on the same test only one variable is measured and a univariate statistic should be used.

Another problem in using the interclass correlation method is that each person is limited to two scores. Thus, in situations where three or more scores are available for each person, the product-moment correlation coefficient is not appropriate unless a modification is made with the resultant reduction of information in the data (Safrit, 1976). This is quite a severe limitation for measures of motor performance since it is very common to administer more than two trials of

a test or to have more than two judges rate each performer.

Split-half method. A second correlational method that has been widely used to estimate reliability is the split-half technique. With this method an even number of trials is administered within one day. Then the test is divided into two parts so that each person has two scores. There are four common methods of splitting a test: (a) sets of items for the two half-tests can be selected on the basis of equivalence in content and difficulty, (b) alternate items (i.e., odd-even) can be placed in each half-test, (c) alternate groups of items can be placed in each half-test, and (d) the first half of the items can be used as one half-test and the second half, as the other half-test (Baumgartner & Jackson, 1975).

For every possible split, a different correlation coefficient is possible for the same set of data. Therefore, the magnitude of the reliability coefficient depends on the method used to divide the scores into two sets. Regardless of the method used to split the test the product-moment correlation between the two sets of scores is "stepped-up" using the Spearman-Brown Prophecy Formula to estimate the reliability of the entire test (Safrit, 1976). Thus, the reliability coefficient for the whole test is interpreted as the correlation coefficient that would be obtained if the whole test was correlated with itself or with another whole test composed of the same number of trials (Baumgartner & Jackson, 1975).

There are several limitations that one should be aware of before using the split-half reliability method. As with the

product-moment correlation, it is a bivariate statistic being used in a univariate situation. Also, variation from trial-to-trial within a half of the test is not considered as error variance in the reliability estimation. Instead, the error variance accounted for is due to change in the score of a subject between the two halves of the test (Safrit, 1976). In addition, Baumgartner (1968) indicates that the split-half coefficient will always be an overestimate of a comparable test-retest reliability coefficient. Finally, Kirkendall, Gruber, and Johnson (1980) emphasize that because of inflated reliability coefficients the split-half and Spearman-Brown Prophecy methods are to be viewed with extreme caution and, in general, be avoided in most physical education situations.

Analysis of variance. The symbol r is used to represent the Pearson product-moment correlation coefficient, an interclass correlation coefficient. When analysis of variance (ANOVA) is used to estimate reliability, the appropriate symbol for the reliability coefficient is R, representing the intraclass coefficient. The coefficient R represents a ratio of variance estimates which provide information on the amount of variance attributable to all measurable sources of variability. By using analysis of variance procedures it is possible to identify specific components of score variation and obtain separate estimates of the relative magnitude of each.

Since reliability estimates for tests of motor performance are generally based on several trials of the test, more than two sets of scores must be taken into account. In a set of measurements

repeated on a group of individuals, the ANOVA method estimates the magnitude of the components of variation and provides a way for determining the consistency with which a variable is measured in a series of repeated testing trials (Brozek & Alexander, 1947). Although all possible inter-trial correlations (r's) could be averaged to obtain a reliability coefficient, use of analysis of variance is more appropriate because a systematic increase or decrease from trial to trial can then be identified as a source of variability (Safrit, 1981).

According to Barrow & McGee (1979), "the analysis of variance technique for establishing reliability has been recommended in the research literature and has merit over the other two methods [Pearson product-moment and split-half]" (p. 40). There are several reasons why the intraclass correlation coefficient appears to be the best method for estimating reliability, whether scores are collected on one day or several days (Baumgartner & Jackson, 1975). As previously mentioned, the ANOVA method permits more than two scores per student. In addition, the intraclass coefficient is sensitive to more sources of measurement error than the other methods, and this gives a truer picture of test reliability (Baumgartner & Jackson, 1975). Finally, the intraclass method is the only method which considers changes in the mean and standard deviation from one set of measures to the next to be measurement error or lack of reliability (Kroll, 1962a).

The alpha coefficient. The alpha coefficient provides an estimate to test reliability equivalent to the intraclass reliability coefficient when a two-way ANOVA is used and the

subjects by trials interaction is used as the error term (Cronbach, 1951). To facilitate the interpretation of alpha, it is important to understand the relationship between the total test variance and the individual trial variances and covariances (Jackson, Jackson, & Bell, 1980). The variance of the total test scores is the sum of all variances and covariances. Thus, the total variance among subjects on a test is determined by both the variance of subjects on each trial and the degree of relationship among trials. Coefficient alpha then, is basically the proportion of total variance that is accounted for by the trial covariances (Jackson, et.al., 1980).

Cronbach (1951) developed the alpha coefficient as an estimate of internal consistency reliability and internal consistency is important if items are viewed as a sample from a relatively homogeneous universe (American Psychological Association, 1974). In addition, Cronbach's alpha coefficient can be applied to multiple-trial motor performance tests because it is applicable to tests where items are scored on a continuous measurement scale (Jackson, et.al., 1980).

RELATED RELIABILITY STUDIES

Of the items selected for this study, some have been examined for reliability on numerous occasions by other researchers. However, as the review will indicate, a good number of the items have not been heavily researched. Only studies conducted on the similar age bracket of 10 to 17 will be reported. The exception to this is where no related studies within the age bracket were found. Then corresponding studies with different ages will be reported.

The studies reported in this study have been evaluated on the following criteria: number, ages, and sex of subjects; method or procedure of testing; number of trials; time between test administrations; and reliability coefficient. If any of the above criteria are not stated, they were not reported in the original study. A summary table can be found in Appendix A.

The single-trial items are presented first in order to separate them from the multi-trial items. Otherwise, the items are discussed in the same order as presented in the Project UNIQUE Training Manual.

Timed Leg Raise

Fleishman (1964a) conducted a study on the timed leg raise in which he tested 201 Navy recruits with the average age of 18 years. He reported a test-retest reliability coefficient of .71.

Timed Trunk Raise

In a study of several physical fitness items, Rarick, Dobbins, and Broadhead (1976) tested 71 males (ages 6 to 9), 65 retarded males (ages 10 to 13), 71 retarded males (ages 6 to 9), 74 females (ages 6 to 9), 61 retarded females (ages 10 to 13), and 64 retarded females (ages 6 to 9). They gave two trials on the same day and reported the following coefficients of reliability: .743 normal boys, .822 old EMR boys, .779 young EMR boys, .705 normal girls, .804 old EMR girls, .843 young EMR girls. For a chest raise done off the end of a table and held 10 inches above the table, Avent (1963) reported a Pearson product-moment correlation of .458 when 50 females,

ages 9 to 12 were tested and retested 3 to 8 weeks apart.

Skinfold Measurements

AAHPERD (1980) reports that test-retest reliability of skinfold fat measures has exceeded .95 in experienced testers. No further information is reported. When testing 162 females and 164 males, ages 10 to 18, with two trials conducted 1 week apart, Colgan (1978) found the following reliability coefficients: triceps--.93 females and .96 males, subscapula--.94 females and .90 males.

Rise To Stand

No studies were found which examined the reliability of the rise to stand test item.

Mat Creep

In conducting three trials on the same day, Rarick, et.al. (1976) found reliability coefficients of: .925 normal boys, .918 old EMR boys, .924 young EMR boys, .940 normal girls, .939 old EMR girls, and .949 young EMR girls.

Shuttle Run

A number of studies have been conducted on the shuttle run. Fleishman (1969b), using a 20-yard distance which was run five times for a total distance of 100 yards, found a test-retest reliability coefficient of .85 after testing 20,000 boys and girls from 45 cities in the United States. Klesius (1968) found within day test-retest reliability of .68 following the testing of 150 tenth-grade males using the AAHPER Youth Fitness Test procedures. Keogh (1965) reported a Pearson product-moment coefficient of .73 for 20 first-grade children and .59 for 24

third-grade children after giving two trials, 2 to 6 weeks apart. After testing 95 junior-high boys and 82 senior-high boys on a 40-yard shuttle run, Baumgartner and Jackson (1970) found intraclass correlation coefficients of .76 and .82 respectively. In a between day test-retest of the same subjects, Baumgartner (1974) found coefficients of .88 for the junior-high boys using the sum of scores on four trials and .93 using the best score. For the senior-high boys he found intraclass coefficients of .87 for the sum of scores and .92 for the best score. Using the AAHPER Youth Fitness Test procedures, Marmis, et.al. (1969) found two trial test-retest reliability coefficients between .60--.80 for 1,122 males, ages 9 to 18 and between .46--.82 for 938 females, ages 9 to 18. From their study, Rarick, et.al. (1976) reported the following test-retest correlations for a shuttle run in which subjects ran a distance of 30 feet four times, for a total distance of 120 feet: .955 normal boys, .915 old EMR boys, .949 young EMR boys, .947 normal girls, .905 old EMR girls, and .926 young EMR girls. Using a 40 foot length which was run four times for a total of 160 feet, Anhalt (1958) reported a Pearson product-moment correlation of .887 after 32 fourth, fifth, and sixth-grade females were given two trials 1 week apart. Also conducting two trials 1 week apart but using the AAHPER Youth Fitness Test procedures, Colgan (1978) found reliability coefficients of .90 for 162 females and .82 for 164 males, ages 10 to 18.

Modified Stork Test

In their administration of the stork test, Rarick, et.al. (1976)

reported the following test-retest coefficients: .790 normal boys, .781 old EMR boys, .791 young EMR boys, .798 normal girls, .727 old EMR girls, and .818 young EMR girls. Subjects were asked to balance on a footboard 3¼ inches high, 3/4 inches wide, and 12 inches long. The authors allowed one practice trial and gave two test trials. In his study of 20,000 boys and girls, Fleishman (1964b) reported a test-retest coefficient of .82. Subjects were asked to balance on a piece of wood 1½ inches high, 3/4 inches wide, and 24 inches long. Two trials were given.

Sit And Reach

For the sit and reach test, AAHPERD (1980) reported that reliability coefficients have been high, ranging above .70. No further information was reported. Giving only one trial after three bobs, Colgan (1978) reported a between day test-retest coefficient of .95 for 162 females and .84 for 164 males, ages 10 to 18.

Grip Strength

On the reliability of grip strength, Fleishman (1964b) reported a reliability coefficient of .91 after testing 20,000 normal subjects ages 13 to 18, giving three trials, only with the preferred hand, with 1 minute rest in between. Rarick, et.al. (1976) gave three trials for each hand while sitting and found test-retest coefficients of: .911 normal boys, .927 old EMR boys, .902 young EMR boys, .882 normal girls, .975 old EMR girls, and .917 young EMR girls for the right grip. For the left grip strength, the same testers (1976) reported coefficients

of .959 normal boys, .941 old EMR boys, .896 normal girls, .959 old EMR girls, and .934 young EMR girls. After testing 23 first graders and 32 third graders, Keogh (1965) found between day (2 to 6 weeks apart) Pearson product-moment test-retest coefficients of: .85 and .75 for the right hand and .79 and .70 for the left hand. He allowed two trials while standing, with 3 seconds rest in between. The same author (1965) found within day Pearson product-moment correlation coefficients of .76 for 25 first graders and .84 for 31 third graders. A between day Pearson product-moment coefficient of .65 for the right hand and .80 for the left hand was reported by Avent (1963), after testing 50 females, ages 9 to 12, with two trials, 3 to 8 weeks apart.

Flexed Arm Hang

In testing three groups of normal children, 1 day apart for Project ACTIVE, Vodola (1978) found Pearson product-moment correlation coefficients of: .97 for 30 females (15 years-old), .88 for 19 females (7 years-old), and .89 for 33 males (15 years-old). Bolonchuk (1971) tested 25 fifth and sixth-grade females using the AAHPER Youth Fitness Test procedures and reported a Pearson product-moment correlation coefficient of .95. Also, using the AAHPER procedures, Colgan (1978) gave two trials, 1 week apart, and reported reliability coefficients of .89 for 162 females and .96 for 164 males, ages 10 to 18. A between day (3 to 8 weeks apart) Pearson product-moment coefficient of .87 was reported by Avent (1963), after testing 50 females, ages 9 to 12. However, an underhand grip was used.

Standing Broad Jump

On the standing broad jump, Klesius (1968) reported a .94 test-retest coefficient following the testing of 50 normal 10th-grade males. Marmis, et.al. (1969) found multi-trial coefficients ranging from .73 to .95 after testing 1,122 boys and 938 girls, ages 9 to 18, on the standing broad jump. Vodola (1978) found test-retest coefficients of .95 for 30 females (15 years-old), .98 for 33 males (15 years-old), .49 for 13 males (6 years-old), and .89 for 19 females (7 years-old). In their study, Rarick, et.al. (1976) reported test-retest coefficients of .805 normal boys, .917 old EMR boys, .947 young EMR boys, .906 normal girls, .953 old EMR girls, and .957 for young EMR girls. After testing 300 males and 300 females, (ages 7, 9, and 11) on twelve trials, Kane and Meredith (1952) found same day Pearson product-moment coefficients between .98 and .99 for the females and between .97 and .99 for the males. The best jump and the second best jump was used to calculate the reliability coefficient. The same authors (1952) gave 12 trials, 2 days apart, to 150 males and females, 7 years-old. Using the best scores from both days, they reported a between day Pearson product-moment correlation of .83 for the males and .86 for the females. Keogh (1965) reported Pearson product-moment coefficients of .90 and .77 following between day testing (2 to 6 weeks apart) of 21 first-grade children and 27 third-grade children. Correlating the best of three trials with the second best trial, the same author (1965) reported a within day Pearson product-moment

coefficient of .91. After testing 95 junior-high boys and 82 senior-high boys on six trials, Baumgartner and Jackson (1970) reported the following ANOVA reliability coefficients: .96 for the junior-high boys using trials one through three and .97 for the senior-high boys using trials three through six. In a between day (1 day apart) study with the same subjects, Baumgartner (1974) found intraclass correlation coefficients of: .96 and .95 for the junior-high and senior-high boys using the sum of scores as the criterion measure. When the best score was used, a coefficient of .96 was reported for both groups. Using the Pearson product-moment method, Bolonchuk (1971) tested and retested 25 males and 25 females. He presented reliability coefficients of .82 for the females and .89 for the males. Avent (1963) reported a Pearson product-moment correlation of .68 for between day test-retest procedures with 50 females (ages 9-12). After testing 32 fourth, fifth, and sixth-grade females, 1 week apart, Anhalt presented a Pearson product-moment reliability coefficient of .913. Finally, Colgan tested 164 males and 162 females, ages 10 to 18, and reported between day (1 week) test-retest coefficients of .82 for females and .81 for males. The AAHPER Youth Fitness Test procedures were used.

Softball Throw (Distance)

For the softball throw, Fleishman (1964b) reported a test-retest reliability of .93 after testing 20,000 normal subjects ages 12 to 18. Klesius (1968) found a coefficient of .93 following the testing of 150 normal 10th-grade males using the AAHPER

Youth Fitness Test procedures. Also, using the AAHPER Youth Fitness Test procedures, Marmis, et.al. (1969) found multi-trial coefficients ranging from .83 to .97 after testing 1,122 boys and 938 girls ages 9 to 18. Three trials were given. After testing 19 first graders and 27 third graders, Keogh (1965) found Pearson product-moment reliability coefficients of .97 and .88 respectively using a 12-inch softball and three trials. Using the best score and the second best score of three trials, the same author (1965) reported a within day Pearson product-moment correlation of .95. Lastly, Bolonchuk (1971) presented reliability coefficients of .93 for 25 females and .94 for 20 males in the fifth and sixth grades. The Pearson product-moment method was used to calculate the coefficients reported.

Softball Throw (Timed)

In their study on the softball throw, Rarick, et.al. (1976) found the following reliability coefficients: .863 normal boys, .964 old EMR boys, .966 young EMR boys, .709 normal girls, .950 old EMR girls, and .854 young EMR girls.

Summary

A review of the literature has revealed the importance of reliability for physical fitness tests. In addition, it has indicated that suitable reliability is often specific to the test and/or population being tested. Furthermore, the importance of developing reliability coefficients directly related to the Project UNIQUE Test Battery and specified populations is apparent. By reviewing the literature, it has become evident that more research is needed on several Project UNIQUE items. They are:

the skinfold measurements, timed leg raise, timed trunk raise, rise to stand, mat creep, modified stork test, sit and reach, and flexed arm hang. In addition, the literature has pointed out the need to further evaluate the shuttle run, grip strength, softball throw and standing broad jump according to Project UNIQUE testing procedures.

CHAPTER III

PROCEDURES

Nature of Information Sought

This study was conducted to determine coefficients of reliability and the standard error of measurement for selected health and performance related fitness test items from the Project UNIQUE Physical Fitness Test Inventory. In addition, the appropriate number of trials for the selected items was determined in order to optimize reliability.

The items selected for this study consisted of both multi-trial items and single-trial items. The multi-trial items were those items on the Project UNIQUE Test Inventory which were administered more than once. Skinfold measurements, rise to stand, mat creep, shuttle run, modified stork test, sit and reach test, grip strength, flexed arm hang, softball throw, and standing broad jump were the multi-trial items being evaluated. Single-trial items were those items on the Project UNIQUE Test which were administered only once. These items were administered once due to the lengthy recovery time that would be needed to insure maximum performance on a repeated trial. The single-trial items evaluated in this study were the timed leg raise and the timed trunk raise.

Sources of Data

Multi-trial items. Subjects included 50 non-impaired youth, 50 visually impaired youth, 50 auditory impaired youth

and 50 orthopedically impaired youth randomly selected from various schools throughout the United States that participated in the Project UNIQUE study. The subjects were between the ages of 10 to 17 and were randomly sampled using a table of random numbers. Permission to test was covered by permission to participate in the Project UNIQUE study.

Single-trial items. Subjects included 50 non-impaired youth, 47 visually impaired youth, and 50 auditory impaired youth randomly selected from at least two different sites for each population studied. Schools selected were from the Buffalo, N.Y. and Rochester, N.Y. areas. The age, sex, and permission conditions were the same as the multi-trial subjects described above.

Procedures for Collecting Data

Multi-trial items. Subjects selected for this study were tested on the following items: skinfold measurements, rise to stand, mat creep, shuttle run, modified stork test, sit and reach test, grip strength, flexed arm hang, standing broad jump, and softball throw. All trials were administered on the same day for the test item measured. However, all test items were not given on the same day. The number of trials per each test item varied according to the Project UNIQUE testing instructions. The number of trials per test item was as follows: (a) skinfold measurements--measurements were taken from the following areas and in the following order: triceps, subscapular, and abdominal (this order was repeated three times so that three measurements were recorded for each area).

(b) rise to stand--three trials, (c) mat creep--three trials, (d) shuttle run--two trials, (e) modified stork test--three trials, (f) sit and reach test--two trials, (g) grip strength--right hand: three trials; left hand: three trials, (h) flexed arm hang-- two trials, (i) standing broad jump--three trials, and (j) softball throw--three trials.

Data was collected during the interval of April 1, 1980 and June 30, 1981. All the test items were administered according to Project UNIQUE testing instructions (see Appendix H). Subjects were tested by certified Project UNIQUE testers only. In order to become a certified Project UNIQUE tester an individual must participate in a 4 hour training program, plus properly demonstrate testing procedures, equipment utilization, and data recording procedures. In addition, certified testers had to successfully pass a written evaluation and receive a minimum score of 90 percent in order to become certified.

The testing is divided into two major areas: (a) demonstrative competencies, and (b) documented competencies. Demonstrative competencies require the trainee to demonstrate the testing procedures and proper equipment utilization for the following tests: grip strength, skinfold measurements, and sit and reach test. Trainees were also required to demonstrate the capacity to accurately record data from measurements taken on sample subjects.

Documented competencies involved the successful completion of a written test. The test was composed of various multiple choice questions that addressed the central purposes, procedures, instructions, modifications, and equipment use for successful

testing (Winnick & Short, 1980). Data transfer capabilities were also documented for trainees.

Single-trial items. Subjects selected for this study were tested on the timed leg raise and timed trunk raise. Single trials were administered on three separate days. Trial days corresponded to three consecutive physical education periods with a fourth period reserved for absentees. Data was collected during the interval of April 1, 1981 and June 30, 1981. All other procedures were the same as for the multi-trial subjects and is described above.

Procedures For Estimating Internal Consistency Reliability

Reliability for multi-trial items is primarily concerned with internal consistency. Internal consistency refers to a consistent rate of student scoring throughout a test or, if multiple trials are given, from trial to trial (Baumgartner & Jackson, 1975). The Kuder-Richardson formula, analysis of variance and coefficient alpha are the primary methods used to determine internal consistency reliability. Since both the intraclass coefficient and alpha coefficient methods are relevant to this study, they will be discussed here.

Intraclass coefficient. According to Safrit (1976), a two-way ANOVA--mixed model should be used when subjects constitute a random group and the trials represent a fixed variable. When this is true, the reliability of the test can be estimated by: $R = \frac{MS_s - MS_{sxt}}{MS_s}$ (Safrit, 1976), where MS_s = mean square for subjects and MS_{sxt} = mean square for subjects

by trials. However, before the above formula can be used a test for trend (randomness of trial means) should be performed to determine the severity of difference among the trial means (Kirkendall, et. al. 1980). To determine whether a test score trend from trial to trial was significant or not the following formula was used: $F = \frac{MS_t}{MS_{sxt}}$ with a significant F indicating trend (Safrit, 1981). If no trend existed then the reliability coefficient was calculated by the above formula recommended by Safrit (1976).

If the test for trend was significant, Kroll (1967) indicated that the mean may not be the most appropriate criterion score. When this was the case, Baumgartner (1969a) recommended examining the trial means and selecting a group of several trials in which trend was not significant. The criterion score then became the mean of the nonsignificant scores and reliability was computed using the formula listed above.

Alpha coefficient. Another method of estimating internal consistency reliability is the alpha coefficient, developed by Cronbach (1951). Cronbach's alpha coefficient is applicable to tests where items are scored on a continuous measurement scale and, for this reason, can be applied to multiple-trial motor performance tests (Jachson, et. al. 1980).

Coefficient alpha is basically the proportion of total variance that is accounted for by the trial covariances. Rather than compute the proportion of covariance directly, Cronbach (1951) utilized the following formula: $\alpha = \frac{k}{k-1}\left(1 - \frac{S_t^2}{S_x^2}\right)$,

where k = total number of trials; S_t^2 = the sum of k trial variances; and S_x^2 = the total test variance (Jackson, et.al., 1980).

The alpha coefficient provides an estimate of test reliability that is equivalent to the intraclass reliability coefficient (see Appendix B) when a two-way ANOVA is used and the subjects by trials interaction is used as the error term (Cronbach 1951). In fact, Jackson, et.al. (1980) recommend combining the intraclass and alpha approaches to obtain the most reliable measurement of a test with multiple-trials. Therefore, this study will combine both methods by using the ANOVA technique to test for trend among trials and using the alpha technique to determine the reliability coefficient.

Analysis of Data

Reliability. The Statistical Packages of the Social Sciences subprogram "Reliability" was used to calculate Cronbach's alpha coefficient of reliability. For the multi-trial items, coefficients of reliability were calculated on the following groups: nonimpaired, visually impaired, auditory impaired, and orthopedically impaired. For the single-trial items, reliability coefficients were determined for the nonimpaired, visually impaired, and auditory impaired subjects. If trend was found among trial means, then reliability was calculated using the most appropriate criterion score (see the section below on determining a criterion score).

For a further analysis of the data, alpha coefficients were computed for younger and older subjects, for all four

groups (nonimpaired, visually impaired, auditory impaired, and orthopedically impaired). In addition, coefficients of reliability were calculated for the subgroups of male and female. However, due to the limited number of subjects for many age and sex subgroups, a test for trend was not performed. These additional coefficients were used for speculation purposes in the discussion section.

Standard error of measurement. The standard error of measurement (SE_m) is an absolute measure of precision (Safrit, 1981). The estimate of the standard error is presented in the actual score units of the data. The intraclass correlation coefficient, on the other hand, is a relative measure of precision because it describes the consistency with which an individual maintains his/her position in the total group when the measurement procedure is repeated. Thus, one advantage the SE_m has over the correlation coefficient is that it is independent of the exact spread of scores (Safrit, 1981).

If it were possible to administer a test repeatedly to an individual, the standard deviation for the distribution of test scores would be the standard error of measurement for that individual (Safrit, 1981). Since it is not practical or often reasonable to administer a test repeatedly, the standard error can be calculated with the following formula: $SE_m = S\sqrt{1 - r}$, where S = the standard deviation and r = the reliability coefficient (Baumgartner & Jackson, 1975).

Determining an appropriate criterion score. To determine an appropriate criterion score the ANOVA technique was used.

Kroll (1967) indicated that "when several trials are available, the assumption of random, uncorrelated error variance due to trials can be tested quite simply by an analysis of variance design for repeated measures" (p. 416). If the between trials F was nonsignificant it was then assumed that trial means did not fluctuate in any pattern. In this case the mean of all trials was used as the criterion measure.

When the between trials F was significant (at the .05 level) a post hoc analysis was done to determine what trials were significant. This process was explained by Dayton (1970), "In the event that the null hypothesis is rejected [a significant F is shown], the experimenter must continue his analysis of the data in order to isolate specific inequalities among the treatment effects" (p. 37).

The Newman-Keuls technique of post hoc analysis was used to determine the significant and nonsignificant trials. For this approach, contrasts of selected series of sequential trials were conducted (Dayton, 1970). When significant differences were located between compared series, the direction of this difference was examined and the trials showing significance were eliminated (Disch, 1975). In addition, this procedure was backed up by analysis of the superdiagonal of the correlation matrix of trials, recommended by Jones, as cited by Disch (1975). In order to locate trend free trials, the superdiagonal was examined to locate the point at which stabilization occurs. Stabilization was defined as the point at which the correlations became equal (Disch, 1975). "Eyeballing" or scanning

the correlation matrix was the technique used to determine stabilization and the trend-free trials.

CHAPTER IV

RESULTS AND DISCUSSION

RESULTS

The purpose of this study was to determine coefficients of reliability and the standard error of measurement for selected health and performance related test items from the Project UNIQUE Physical Fitness Test Inventory. In addition, the appropriate number of trials and correct criterion score were determined by an analysis of the superdiagonal elements of the correlation matrix.

Subjects included 50 nonimpaired youth, 50 visually impaired youth, 50 auditory impaired youth, and 50 orthopedically impaired youth between the ages of 10 and 17 randomly selected from the various schools participating in the Project UNIQUE study. These subjects were measured on selected health and physical fitness items from the Project UNIQUE Test Inventory. Three trials were administered on the following items: skinfold measurements, rise to stand, mat creep, modified stork test, grip strength, standing broad jump, and softball throw. Two trials were administered for the shuttle run, sit and reach, and flexed arm hang test items.

For the single-trial items, subjects included 50 nonimpaired youth, 47 visually impaired youth, and 50 auditory impaired youth between the ages of 10 and 17. Subjects were randomly selected from at least two different sites for each population tested. Sites were located in the Rochester, N.Y.

and Buffalo, N.Y. areas. These subjects were tested for among day reliability and given three trials on both the timed leg raise and timed trunk raise. Each trial was administered on three consecutive physical education periods.

Reliability coefficients were determined by using the SPSS subprogram to calculate Cronbach's alpha coefficient. To determine a criterion measure an analysis of variance was calculated with a significant F indicating trend. When a between trials trend was shown, a post hoc analysis was performed to determine which trials were significantly different. The Newman-Keuls technique of post hoc analysis was used. Decisions concerning appropriate criterion trials were collaborated by inspection of the superdiagonal pattern of the correlation matrix.

The results being reported in this section include the alpha coefficient, the standard error of measurement, the recommended criterion score(s) and the mean of the criterion score(s). It should be noted that this data was calculated using trend free trials. A summary of this information is presented in Tables 1-6. For the data calculated using all the trials, the reader is referred to Appendix C.

It must be noted that the reliability of a test or instrument is dependent on the type of measure, age and sex of the respondents, abilities of the administrator, and so on, making it impossible to specify a minimum acceptable reliability (Baumgartner, 1975). According to Kelly (1927), the

evaluation of the size of reliability cofficients should be made in light of the types of decisions the test user will make based on the test results. For example, where decisions of group accomplishment are concerned a value of .50 may be adequate. However, if the individual is concerned with the level of individual accomplishment (such as using the score to determine a grade) a minimal reliability coefficient of .94 is recommended.

For discussion purposes, this study will use the following labels and cutoff points for its reliability coefficients: .90 and above--excellent or high, .89-.80 very good, .79-.50 moderate or fair, .49-.20 low or weak, and .19 and below--very poor. It must be emphasized that these are rough cutoff points and are only intended to be used for the current data.

Timed Leg Raise

A test for independence of trial means resulted in a significant $F = 4.17$, at $p = .018$, for nonimpaired youth. The post hoc analysis revealed that only trials two and three were not significantly different from each other and they also had the highest inter-item correlation ($r = .87$). Therefore, the mean of trials two and three was selected as the criterion score. The mean (M) of the nonsignificant trials was 42.89 secs. and an alpha reliability coefficient (α) of .93 was determined using trials two and three. The standard error of measurement (SE_m) was 6.38.

For the visually impaired, a significant $F = 3.89$ was obtained, $p = .023$. Trials one and three were determined to

be significantly different by post hoc analysis. It was decided to use the mean of trials two and three as the criterion measure as they had the highest inter-item correlation (r = .93) and were not significantly different. The mean of the two trials was 44.65 secs., the reliability coefficient was .94 and 11.34 was the SE_m.

The mean of all trials was selected as the criterion score for auditory impaired as a nonsignificant F was determined at $p > .05$ level. The mean was 53.93 secs. with an alpha coefficient of .93 and a SE_m of 9.12. The results of the timed leg raise are presented in Table 1.

Timed Trunk Raise

A significant F = 4.04 was reported at p = .020 for the nonimpaired subjects. It was decided to use trial two as the criterion measure as post hoc analysis revealed trials one and two to be signficantly different, while trials two and three and one and three had the lowest inter-item correlations. The mean of trial two reported to be 68.06 secs. An alpha reliability coefficient of .85 was determined by using trials two and three, as they were nonsignificant. The SE_m was 21.81 using the standard deviation of the criterion score and the alpha coefficient of trials two and three.

A nonsignificant F at the $p > .05$ level was reported for the visually impaired. Using the mean of all three trials as the criterion score the following results were calculated: M = 47.75 secs., α = .99, and SE_m = 4.57.

For the auditory impaired, a nonsignificant F, $p > .05$,

was calculated. Therefore, the mean of all three trials was selected as the criterion score and the following results were determined: $\underline{M}$ = 59.37 secs., α = .91, and $\underline{SE}_m$ = 11.40. The results of the timed trunk raise are presented in Table 1.

Table 1

Summary Data on Leg Raise and Trunk Raise

Calculated Using Trend Free Trials

Variable	Trials	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$
		Nonimpaired				Visual			
L Raise*	3	$\underline{M}_{2,3}$	42.89	.93	6.38	$\underline{M}_{2,3}$ 47[1]	44.65	.94	10.16
T Raise*	3	2	68.06	.85	21.81	$\underline{M}_{1,2,3}$ 47[1]	47.75	.99	4.57
		Auditory				Orthopedic			
L Raise*	3	$\underline{M}_{1,2,3}$	53.93	.93	9.12	---	---	---	---
T Raise*	3	$\underline{M}_{1,2,3}$	59.37	.91	11.40	---	---	---	---

*single trial among days

[1]number of subjects

Skinfold Measurements

Triceps. At the $p > .05$ level a nonisgnificant $\underline{F}$ was obtained for the nonimpaired youth. The mean of 12.72 mms. for all three trials, was selected as the criterion measure and the

results were: α = .99 and SE_m = .33.

The mean of all three trials was also selected as the criterion score for the visually impaired as a nonsignificant F was calculated. The following results were then determined: M = 13.75 mms., α = .97, and SE_m = 1.04.

For the auditory impaired, a nonsignificant F at $p > .05$ was achieved causing the mean of all three trials to be selected for the criterion measure. The calculated results were: M = 12.49 mms., α = .98, and SE_m = .79.

As with the three prior groups, a nonsignificant F was also reported for the orthopedically impaired. Thus, the mean of all three trials (12.70 mms.) was selected as the criterion score, with the following results: α = .97 and SE_m = .85. The results of the triceps measurements are presented in Table 2.

Abdominal. The mean of all three trials was selected as the criterion score as a nonsignificant F was achieved for the nonimpaired group. The reported results were: M = 14.14 mms., α = .99, and SE_m = .54.

For the visually impaired, a nonsignificant F at $p > .05$ was also calculated. Therefore, the mean of all three trials (16.68 mms.) was chosen as the criterion measure and the following results calculated: α = .90 and SE_m = 2.66.

A significant F = 5.28 was reported at the $p < .006$ level for the auditory impaired. Post hoc analysis revealed trials one and two and trials one and three to be significantly different. Therefore, the mean of trials two and three was chosen as the criterion score. The following results were then calculated: M = 13.45 mms., α = .99, and SE_m = .77.

Forty-nine orthopedically impaired youth were tested on the abdominal skinfold. The ANOVA summary table reported a non-significant F at $p > .05$. Using the mean of all three trials as the criterion score the following results were calculated: M = 13.93 mms., α = .99, and SE_m = .53. The results of the abdominal measurements are presented in Table 2.

Subscapula. The mean of all three trials was selected as the criterion measure for the nonimpaired group, as a non-significant F was calculated. The following results were then determined: M = 10.29 mms., α = .99, and SE_m = .36.

The visually impaired also achieved a nonsignificant F at $p > .05$ and the mean of all three trials was also selected as the criterion score. The reported results were: M = 13.84 mms., α = .99, and SE_m = .60.

For the auditory impaired the ANOVA summary table revealed a significant F = 5.19 at p = .007. After performing a post hoc analysis, it was discovered that trials one and three and two and three were significantly different, while trials one and two were not. Therefore, the mean of trials one and two was chosen as the criterion measure and used to calculate the reliability coefficient and standard error of measurement. The following results were determined: M = 10.30 mms., α = .99, and SE_m = .57.

A nonsignificant F resulted for the orthopedically impaired causing the mean of all three trials to be chosen as the criterion measure. The reported results were: M = 10.64 mms., α = .99, and SE_m = .35. The results of the subscapula

measurements are presented in Table 2.

Table 2

Summary Data on Skinfold Measurements

Calculated Using Trend Free Trials

Variable	Trials	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$
		Nonimpaired				Visual			
Triceps	3	$\underline{M}_{1,2,3}$	12.72	.99	.33	$\underline{M}_{1,2,3}$	13.75	.97	1.04
Abdom	3	$\underline{M}_{1,2,3}$	14.14	.99	.54	$\underline{M}_{1,2,3}$	16.68	.90	2.66
Subscp	3	$\underline{M}_{1,2,3}$	10.29	.99	.36	$\underline{M}_{1,2,3}$	13.84	.99	.60
		Auditory				Orthopedic			
Triceps	3	$\underline{M}_{1,2,3}$	12.49	.98	.79	$\underline{M}_{1,2,3}$	12.70	.97	.85
Abdom	3	$\underline{M}_{2,3}$	13.45	.99	.77	$\underline{M}_{1,2,3}$	13.93	.99	.53
Subscp	3	$\underline{M}_{1,2}$	10.30	.99	.57	$\underline{M}_{1,2,3}$	10.64	.99	.35

[1]number of subjects 49

Rise To Stand

For the rise to stand item, a nonsignificant $\underline{F}$ at $\underline{p} > .05$ was found for all four groups. Therefore, the mean of all three trials was selected as the criterion score for each group. The results for each group follow: nonimpaired: $\underline{M}$ = 1.38 secs., α = .40, $\underline{SE}_m$ = .38; visually impaired: $\underline{M}$ = 1.88 secs., α = .91, $\underline{SE}_m$ = .22; auditory impaired: $\underline{M}$ = 1.67 secs., α = .90,

SE_m = .16; and orthopedically impaired: M = 4.39 secs., α = .93, and SE_m = .51 (25 subjects were tested). The results of the rise to stand are presented in Table 3.

Mat Creep

The ANOVA summary table reported a significant F = 3.48 at p = .034 for nonimpaired subjects. After conducting a post hoc analysis, it was discovered that trials one and three were significantly different and that trials two and three had the highest inter-item correlation (r = .67) of the nonsignificant trials. Thus, the mean of trials two and three was chosen as the criterion score. Using trials two and three the following results were reported: M = 3.51 secs., α = .80, SE_m = .18.

For visually impaired, auditory impaired, and orthopedically impaired a nonsignificant F at $p > .05$ was calculated and the mean of all three trials was chosen as the criterion measure. The reported results were: visually impaired: M = 4.42 secs., α = .97, and SE_m = .23; auditory impaired: M = 3.86 secs., α = .96, and SE_m = .20; orthopedically impaired: M = 10.23 secs., α = .97, and SE_m = 1.22 (41 subjects were tested). The results for the mat creep are presented in Table 3.

Shuttle Run

A significant F = 5.57 at p = .022 was reported for the nonimpaired group. Due to the fact that there were only two trials, the mean of trial two was selected as the criterion score because it represented the best score. The reliability coefficient and standard error of measurement were calculated

using trials one and two because only two trials were given. The results reported were: $\underline{M}$ = 11.16 secs., α = .86, and $\underline{SE}_m$ = .47.

A nonsignificant $\underline{F}$ was determined for the visually impaired, auditory impaired, and orthopedically impaired. The mean of both trials was selected as the criterion measure for all three groups. The calculated results of each group were: visually impaired: $\underline{M}$ = 12.49 secs., α = .92, $\underline{SE}_m$ = .68; auditory impaired: $\underline{M}$ = 11.95 secs., α = .70, $\underline{SE}_m$ = 2.48; and orthopedically impaired: $\underline{M}$ = 34.77 secs., α = .99, $\underline{SE}_m$ = 2.04. The results for the shuttle run are presented in Table 3.

Table 3

Summary Data on the Rise to Stand, Mat Creep, and Shuttle Run Calculated Using Trend Free Trials

Variable	Trials	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$
		Nonimpaired				Visual			
Stand	3	$\underline{M}_{1,2,3}$	1.38	.40	.38	$\underline{M}_{1,2,3}$	1.88	.91	.22
M Creep	3	$\underline{M}_{2,3}$	3.51	.80	.18	$\underline{M}_{1,2,3}$	4.42	.97	.23
SR Time	2	2	11.16	.86	.47	$\underline{M}_{1,2}$	12.49	.92	.68
		Auditory				Orthopedic			
Stand	3	$\underline{M}_{1,2,3}$	1.67	.90	.16	$\underline{M}_{1,2,3}^{25}$[1]	4.39	.93	.51
M Creep	3	$\underline{M}_{1.2.3}$	3.86	.96	.20	$\underline{M}_{1,2,3}^{41}$[1]	10.23	.97	1.22
SR Time	2	$\underline{M}_{1,2}$	11.95	.70	2.48	$\underline{M}_{1,2}$	34.77	.99	2.04

[1]number of subjects

Modified Stork Test

A nonsignificant F at $p > .05$ was reported for all four groups on the modified stork test. The mean of all three trials was then selected as the criterion measure for all four groups. Listed below are the results. Nonimpaired: M = 46.96 secs., α = .89, SE_m = 26.92; visually impaired: M = 17.15 secs., α = .77, SE_m = 10.06; auditory impaired: M = 9.89 secs., α = .76, SE_m = 7.52; and orthopedically impaired: M = 2.42 secs., α = .61, SE_m = 2.86 (twenty-two subjects were tested). The results of the stork test are presented in Table 4.

Sit And Reach

The ANOVA summary tables reported a significant F for all four groups on the sit and reach. In addition, a post hoc analysis revealed trials one and two to be significantly different for all four groups. Therefore, the best score (trial two) was selected as the criterion measure for all four groups. However, trials one and two were used to calculate the alpha coefficient and the standard error of measurement **because only two trials were given.**

The following results were reported for the nonimpaired: F = 35.32 at $p < .001$, M = 29.90 cms., α = .99, and SE_m = .93. For the visually impaired the results were: F = 35.32 at $p < .001$, M = 24.18 cms., α = .98, and SE_m = 1.14. Results for the auditory impaired were: F = 12.53 at $p < .001$, M = 23.42 cms., α = .99, and SE_m = .94. For the orthopedically impaired the following results were determined: F = 13.17 at

$p < .001$, $\underline{M}$ = 21.52 cms., α = .99, and $\underline{SE}_m$ = .97 (38 subjects were tested). The results of the sit and reach test are presented in Table 4.

Table 4

Summary Data on the Stork Test and Sit and Reach Calculated Using Trend Free Trials

Variable	Trials	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$
		Nonimpaired				Visual			
Stork	3	$\underline{M}_{1,2,3}$	46.96	.89	26.92	$\underline{M}_{1,2,3}$	17.15	.77	10.06
S Reach	2	2	29.90	.99	.93	2	24.18	.98	1.14
		Auditory				Orthopedic			
Stork	3	$\underline{M}_{1,2,3}$	9.89	.76	7.52	$\underline{M}_{1,2,3}^{22[1]}$	2.42	.61	2.86
S Reach	2	2	23.42	.99	.94	$2^{38[1]}$	21.52	.99	.97

[1]number of subjects

Grip Strength

Right grip. A significant $\underline{F}$ = 8.18 at $p < .001$ was found for the nonimpaired group. The post hoc analysis revealed trials one and two and trials one and three to be significantly different. However, since the trial means (24.8 kgs., 23.7 kgs., and 23.4 kgs.) had a decreasing value it was decided to select trial one as the criterion score. Trials one and two were

then used to calculate the reliability coefficient and the standard error of measurement. The following results were obtained: $\underline{M}$ = 24.84 kgs., α = .96, and $\underline{SE}_m$ = 1.42.

For the visually impaired, a nonsignificant $\underline{F}$ at $\underline{p} > .05$ was computed. Thus, the mean of all three trials was chosen as the criterion score. The results were: $\underline{M}$ = 23.58 kgs., α = .99, and $\underline{SE}_m$ = 1.14.

The ANOVA summary table showed a significant $\underline{F}$ = 5.32 at $\underline{p}$ = .006 for the auditory impaired. A post hoc analysis revealed trials one and three and trials two and three to be significantly different. Trials one and two were not significantly different, thus, their mean score was chosen as the criterion measure. In addition, the reliability coefficient and standard error of measurement were determined by using trials one and two. The criterion score had the following results: $\underline{M}$ = 23.85 kgs., α = .95, and $\underline{SE}_m$ = 1.84.

The mean of all three trials was selected as the criterion score for the orthopedically impaired group because a nonsignificant $\underline{F}$ was reported. The following results were determined: $\underline{M}$ = 14.15 kgs., α = .98, and $\underline{SE}_m$ = 1.33. The results of the right grip strength are presented in Table 5.

Left grip. A significant $\underline{F}$ = 10.14 at $\underline{p} < .001$ was found for the nonimpaired youth. In addition, a post hoc analysis revealed that all three trials were significantly different from each other. Due to the decreasing values of the trial means (23.2 kgs., 22.3 kgs., and 21.6 kgs.) it was decided to use trial one as the criterion score, while trials one

and two were used to calculate the reliability coefficient and standard error of measurement. The calculated results were: $\underline{M}$ = 23.20 kgs., α = .98, and $\underline{SE}_m$ = 1.27.

The $\underline{F}$ statistic was reported as nonsignificant at $\underline{p} > .05$, for the visually impaired, and the mean of all three trials was used as the criterion score. The following results were obtained: $\underline{M}$ = 21.17 kgs., α = .97, $\underline{SE}_m$ = 1.69.

For the auditory impaired, a significant $\underline{F}$ = 3.21 at $\underline{p}$ = .045 was discovered. Trials one and three were determined to be significantly different by post hoc analysis. Since the trial means had a decreasing effect (21.6 kgs., 20.9 kgs., and 20.5 kgs.) and because trial one and two had the highest inter-item correlation ($\underline{r}$ = .96), the mean of trials one and two were selected as the criterion score. The obtained results were: $\underline{M}$ = 21.25 kgs., α = .98, and $\underline{SE}_m$ = 1.26.

The ANOVA summary table showed a significant $\underline{F}$ = 9.55 at $\underline{p} < .001$, for the orthopedically impaired youth. An analysis revealed trials one and three and trials two and three to be significantly different. Therefore, the mean of trials one and two was selected as the criterion score and used to calculate the reliability coefficient and standard error of measurement. Results were: $\underline{M}$ = 15.29 kgs., α = .99, and $\underline{SE}_m$ = 1.19. The results of the left grip strength are presented in Table 5.

Flexed Arm Hang

A nonsignificant $\underline{F}$ at $\underline{p} > .05$ was reported for all four groups on the flexed arm hang. As a result, the mean of both trials was selected as the criterion score. The following

results were obtained for each group: nonimpaired: $\underline{M}$ = 8.35 secs., α = .93, $\underline{SE}_m$ = 2.88; visually impaired: $\underline{M}$ = 11.48 secs., α = .84, $\underline{SE}_m$ = 4.85; auditory impaired: $\underline{M}$ = 9.17 secs., α = .96, $\underline{SE}_m$ = 3.11; and, orthopedically impaired: $\underline{M}$ = 2.50 secs., α = .96, $\underline{SE}_m$ = 1.08. The results of the flexed arm hang are presented in Table 5.

Table 5

Summary Data on Grip Strength and Arm Hang

Calculated Using Trend Free Trials

Variable	Trials	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$	Crit. score(s)	$\underline{M}$	α coef.	$\underline{SE}_m$
		Nonimpaired				Visual			
R Grip	3	1	24.84	.96	1.42	$\underline{M}_{1,2,3}$	23.58	.99	1.14
L Grip	3	1	23.20	.98	1.27	$\underline{M}_{1,2,3}$	21.17	.97	1.69
Arm Hang	2	$\underline{M}_{1,2}$	8.35	.93	2.88	$\underline{M}_{1,2}$	11.48	.84	4.85
		Auditory				Orthopedic			
R Grip	3	$\underline{M}_{1,2}$	23.85	.95	1.84	$\underline{M}_{1,2,3}$	14.15	.98	1.33
L Grip	3	$\underline{M}_{1,2}$	21.25	.98	1.26	$\underline{M}_{1,2}$	15.29	.99	1.19
Arm Hang	2	$\underline{M}_{1,2}$	9.17	.96	3.11	$\underline{M}_{1,2}$	2.50	.96	1.08

Standing Broad Jump

The mean of all three trials was chosen as the criterion measure because a nonsignificant $\underline{F}$ was found for the nonimpaired

students. The obtained results follow: $\underline{M}$ = 5.27 ft., α = .96, and $\underline{SE}_m$ = .18.

A significant $\underline{F}$ = 3.57 at $\underline{p}$ = .032 was reported for the visually impaired. Post hoc analysis revealed that trials one and three were significantly different and showed trials two and three to have the highest inter-item correlation ($\underline{r}$ = .93). In addition, the trial means increased monotonically (4.8 ft., 5.0 ft., and 5.1 ft.). Therefore, it was decided to use the mean of trials two and three as the criterion measure, as well as for calculating the reliability coefficient and standard error of measurement. The following results were reported: $\underline{M}$ = 5.03 ft., α = .97, and $\underline{SE}_m$ = .23.

For both the auditory impaired and the orthopedically impaired, a nonsignificant $\underline{F}$ was reported. Therefore, the mean of all three trials was chosen as the criterion measure. The results for the auditory impaired were: $\underline{M}$ = 5.08 ft., α = .98, and $\underline{SE}_m$ = .16. For the orthopedically impaired (21 subjects were tested) the results were: $\underline{M}$ = 2.04 ft., α = .99, and $\underline{SE}_m$ = .16. The results of the standing broad jump are presented in Table 6.

Softball Throw (Time)

For all four groups a nonsignificant $\underline{F}$ at $\underline{p} > .05$ was found on the softball throw for time. Thus, the mean of all three trials was selected as the criterion measure. The following results were obtained for each group: nonimpaired: $\underline{M}$ = 2.13 secs., α = .84, $\underline{SE}_m$ = .24; visually impaired: $\underline{M}$ = 1.88 secs., α = .88, $\underline{SE}_m$ = .26; auditory impaired: $\underline{M}$ =

2.03 secs., α = .88, SE_m = .22; and orthopedically impaired (49 subjects): M = 1.05 secs., α = .94, SE_m = .14. The results of the softball throw (time) are presented in Table 6.

Softball Throw (Distance)

A nonsignificant F was reported on all four groups, therefore, the mean of all three trials was selected as the criterion score. The results for each group were: nonimpaired: M = 92.39 ft., α = .95, SE_m = 7.63; visually impaired: M = 70.71 ft., α = .99, SE_m = 4.35; auditory impaired: M = 84.73 ft., α = .99, SE_m = 4.11; and orthopedically impaired: M = 28.91 ft., α = .87, and SE_m = 4.51. The results of the softball throw (distance) are presented in Table 6.

Table 6

Summary Data on Broad Jump, Softball Throw (Time) and Softball Throw (Distance) Calculated Using Trend Free Trials

Variable	Trials	Crit. score(s)	M	α coef.	SE_m	Crit. score(s)	M	α coef.	SE_m
		Nonimpaired				Visual			
B Jump	3	$M_{1,2,3}$	5.27	.96	.18	$M_{2,3}$	5.03	.97	.23
SB Time	3	$M_{1,2,3}$	2.13	.84	.24	$M_{1,2,3}$	1.88	.88	.26
SB Dist	3	$M_{1,2,3}$	92.39	.95	7.63	$M_{1,2,3}$	70.71	.99	4.35
		Auditory				Orthopedic			
B Jump	3	$M_{1,2,3}$	5.08	.98	.16	$M_{1,2,3}^{21}$[1]	2.04	.99	.16
SB Time	3	$M_{1,2,3}$	2.03	.88	.22	$M_{1,2,3}^{49}$[1]	1.05	.94	.14
SB Dist	3	$M_{1,2,3}$	84.73	.99	4.11	$M_{1,2,3}$	28.91	.87	4.51

[1] number of subjects

Further Analysis

For a further analysis, the data was subdivided into the following categories: younger (ages 10 to 13), older (ages 14 to 17), males and females. A summary of this data is presented in Appendices D--G. This section of the paper will only report unusually low alpha coefficients or coefficients significantly lower than that achieved by the four sample groups.

On the rise to stand test items, an alpha coefficient of .37 was determined for the younger nonimpaired subgroup and for nonimpaired females. This coefficient is slightly lower than the .40 achieved by all nonimpaired subjects. In addition, a moderate coefficient of .79 was achieved by the males; while a strong coefficient of .88 was reported for the older nonimpaired group.

On the shuttle run, nonimpaired males achieved an alpha coefficient of .79, appreciably lower than the other three subgroups (.83, .93, and .89). For the stork test, the older nonimpaired subgroup had a coefficient of .67, which is below acceptable levels. Also, an alpha coefficient of .73 was reported for nonimpaired males, which was well below the .89 reported for the younger and female nonimpaired subgroups.

Alpha coefficients of .77 and .70 were determined, on the softball throw for time, for the subgroups of younger nonimpaired and nonimpaired males. While higher coefficients of .91 (old nonimpaired) and .87 (females) were determined for the other two subcategories.

A modest coefficient of .77 was reported on the stork test for the visually impaired group. This is reflected in the coefficients determined for the older, male, and female subcategories (.74, .78, and .75), but not reflected in the coefficient of .86 reported for the younger visually impaired.

Although the alpha coefficients on the flexed arm hang for older and male visually impaired subgroups might meet acceptable reliability standards (.80 & .81), they are lower than those determined for the younger and female subcategories (.90 & .88).

For the auditory impaired subgroups of older and females on the shuttle run, alpha coefficients of .70 and .69 were determined, while coefficients of .89 and .80 were reported for the younger and male subcategories.

On the stork test a low coefficient of .59 was calculated for the subgroup of auditory impaired males, and a fair coefficient of .72 for the older auditory impaired subjects. The former coefficient is substantially lower than the .81 reported for both the younger and female subcategories.

An alpha coefficient of .64 was determined for the auditory impaired females on the arm hang. This was substantially lower than the other three groups (.90, .87 & .93).

For the orthopedically impaired on the stork test, a modest alpha coefficient of .76 was reported for the younger subgroup and an unacceptable coefficient of .61 for the older subgroup. However, a high coefficient of .90 was determined for orthopedically impaired females. Unfortunately, there

were insufficient numbers to report a reliability coefficient for the male subgroup.

Although the alpha coefficients on the softball throw for older and female orthopedically impaired (.82 & .81) may meet acceptable standards, they are noticeably below the .93 determined for the younger orthopedically impaired subjects and well below the .98 reported for the male subgroup.

DISCUSSION

Timed Leg Raise

Three trials, on three different days, were administered for the timed leg raise and the mean of all three trials was initially used as the criterion score. However, results indicated that for the nonimpaired and visually impaired groups the mean of trials two and three would be a more appropriate criterion score. Using trials two and three a higher alpha coefficient (.93 & .94 vs. .83 &.91) and lower standard error of measurement (6.4 & 10.1 vs. 9.5 & 11.3) was achieved for both groups. Thus, trial one should serve as a warm-up or practice trial for these groups. This is especially true for the subgroups of younger, male, and female nonimpaired subjects, as a further analysis showed the mean of trials two and three to be a more appropriate criterion measure.

For the auditory impaired, three trials should still be given with the mean of all three trials representing the criterion score. It is unclear why trial one was not significantly different for the auditory impaired group but was

significantly different for the other groups.

The high reliability coefficients achieved when the non-significant trials were used indicate that the test item can be reliable. Coefficients for this study (.93 & .94) were substantially higher than the coefficient (.71) reported by Fleishman (1964a), in the only other study on the timed leg raise. Therefore, it is recommended that three trials continue to be given with the first trial not being scored for nonimpaired and visually impaired subjects.

Timed Trunk Raise

For the timed trunk raise, three trials on three separate days were given. The trials proved not to be significantly different and showed strong reliability (.99 & .91) for the visually and auditory impaired. However, for the nonimpaired group, trials one and two were determined to be significantly different. By examining the means of trials two and three it was determined that trial two would be the most appropriate score and that trial one should not be counted.

It is unclear why such a difference among groups occurred, however, one possible explanation may be the motivational factor. When testing the visually and auditory impaired subjects it was obvious that they were highly motivated and tried their best to improve upon previous scores. This was not the case for the nonimpaired subjects who were anxious to finish the testing and return to their physical education class, which was already in progress.

The reliability coefficients achieved in this study

(.85, .99 & .91) were higher than similar studies. Giving one trial on two different days, 3 to 8 weeks apart, Avent (1963) reported a low reliability coefficient of .46. Perhaps the length of time between trials caused the lower reliability coefficient as Safrit (1981) points out that the time between testing should not be long. Using within day procedures, Rarick, et.al. (1976) reported reliability coefficients ranging from .71 to .84, giving two trials. Thus, it may be more reliable to use between day testing if more than one trial is given. This may be due to a muscular fatigue factor, as Fleishman (1964a) indicates that these items (dynamic strength/endurance) require an all out muscular effort with a progressive decrement in force. Therefore, same day trials may not be accurate measurements unless adequate rest is given between trials.

Skinfold Measurements

The alpha coefficients achieved for the skinfold measurements were very high ranging between .97--.99, with one coefficient of .90. These coefficients exceeded those reported by Colgan (1978) in the only other study on similar age groups (.90--.96). In addition, these coefficients equaled or exceeded the reliability coefficients reported by others (Jackson, et.al., 1978; Wilmore & Behnke, 1969, 1970) on older subjects. Therefore, the procedures used to train Project UNIQUE testers resulted in highly reliable data. This was verified by Sinning (1980) who indicated that skinfold measurements are usually highly reliable when testers are allowed to practice the

procedures before taking measurements. In addition, Sinning pointed out that marking the sites will increase reliability.

It should be noted that in comparing the results among the groups tested, it was discovered that for the auditory impaired a more appropriate criterion score would be the mean of trials two and three for abdominal skinfolds and the mean of trials one and two for subscapula skinfolds. This writer does not recommend changing the testing procedures currently being used, however, these results showed the possibility of error when using a large number of testers. For greater accuracy, it would be wiser to mark the measurement sites as recommended by Sinning (1980).

Rise To Stand

The alpha coefficients for the rise to stand test item were high (.90--.93) for the visually, auditory, and orthopedically impaired groups. However, a low coefficient of .40 was determined for the nonimpaired group. A further analysis revealed a low coefficient of .37 for the younger nonimpaired subjects and nonimpaired females. For the older nonimpaired, a very good coefficient of .88 was determined, while a moderate coefficient of .79 was reported for nonimpaired males. These results indicated that the low reliability for the nonimpaired group was caused by the younger and female subjects.

After examining the trial means of these subjects (younger--1.36 secs., 1.34 secs., and 1.49 secs.; females--1.36 secs., 1.30 secs., and 1.48 secs.) it was evident that trial three was the worst trial. This was not the case for the older and

male subgroups (older--1.38 secs., 1.28 secs., and 1.32 secs.; males--1.39 secs., 1.39 secs., 1.38 secs.) for whom trial three was either the best or second best trial. It is possible that boredom and/or fatigue may have caused trial three to be the worst trial for the younger and female subjects thereby lowering the reliability for these subgroups. However, further research is needed to determine the exact cause of the lower reliability.

Another possibility for the lower reliability for the nonimpaired group is the variability among subjects. An examination of the standard deviations for the nonimpaired shows a large difference between the first two trials and trial three (.25 & .26 to 1.03). This inconsistency is not evident in the other groups and may have caused the lower reliability (Safrit, 1981).

Due to the fact that trial means were not significantly different for all four groups, this study would confirm the Project UNIQUE procedure of giving three trials and using the mean of all three trials as the criterion score.

Mat Creep

Very high alpha coefficients (.96--.97) were determined for the visually, auditory, and orthopedically impaired groups, showing that the mat creep was highly reliable for the impaired populations. A moderate coefficient of .80 was reported for the nonimpaired group. By analyzing the alpha coefficients for the nonimpaired subgroups (.82 younger, .90 older, .89 males, and .84 females), it appears that the

problem lies with the younger nonimpaired subjects. After examining the trial means for this subgroup (3.57 secs., 3.55 secs., and 3.45 secs.) it was evident that the younger subjects improved with each trial. This improvement indicated a possible learning effect which may have caused the lower reliability. However, additional research is needed in order to substantiate this finding.

The results for the mat creep also indicated that three trials be given and the mean of all three trials be used as the criterion measure. This was true for all the impaired groups, however, for the nonimpaired subjects the results indicated that the mean of trials two and three was a more appropriate measure.

Shuttle Run

The Project UNIQUE Training Manual recommends two trials for the shuttle run with the mean of both trials being used to represent test performance. Results from this study confirm this procedure for the impaired populations. The results also indicate that trial two was the appropriate criterion score for the nonimpaired students. Perhaps a practice trial is needed and trial one should be omitted from the scoring procedure.

The alpha coefficients for the nonimpaired, visually impaired, and orthopedically impaired groups were very good to very high (.86, .92, and .99). These coefficients were better than the reliability coefficients reported in prior studies (see Appendix A). Only the study by Rarick, et. al.

(1976) reported comparable coefficients (.91 to .96).

A fair coefficient of .70 was reported for the auditory impaired. This fair coefficient is reflected in the subgroups of older auditory impaired and auditory impaired females, for which coefficients of .70 and .69 were determined. The fair reliability coefficients may be explained by the fact that girls either level off or decrease in motor ability during adolescence (Fleishman, 1964a & Rarick, 1973). In addition, deaf children have been found inferior on tests involving locomotor coordination (Sherrill, 1977). Therefore, lower skilled subjects normally are less consistent in their scoring and can have a lower reliability (Safrit, 1981). This inconsistent performance can cause a leveling off of performance which can cause minor variations in scores to alter the relative order of performers thus reducing reliability. In addition, Kroll (1970) indicated that lower skilled females have displayed less error variance than higher skilled females in tests involving maximal efforts. Futhermore, these lower skilled subjects had a lower true score variance than the highly skilled subjects, and as a result, reflected less reliability.

Modified Stork Test

For the stork test, the Project UNIQUE Training Manual recommends that three trials be given with the mean of all three trials representing the test score. The results of this study confirm this procedure for all four groups.

An alpha coefficient of .89 was calculated for the

nonimpaired subjects. This coefficient proved to be higher than the coefficients of reliability reported by Fleishman (1964b) [.82] and Rarick, et. al. (1976) [.72 to .82], in the only other related studies. Thus the stork test appears to be a reliable item for the nonimpaired group. However, a fair coefficient (.67) was reported for the older subgroup. This may not be an accurate statistic due to the small number (13) of older nonimpaired subjects used in the study. Nevertheless, it indicates a possible cause of concern.

Moderate alpha coefficients of .77, .76, and .61 for the visually, auditory, and orthopedically impaired students indicate that the stork test may not be reliable for these populations. The coefficients of .77 and .76 compare well to the coefficients (.73--.82) reported by Rarick, et. al. (1976), while the alpha coefficient of .61 does not.

In analyzing the reliability coefficients of the visually impaired subgroups for further information, fair coefficients were reported for the older, male, and female subgroups (.74, .78, and .75). These results confirm the questionable reliability of the stork stand for visually impaired subjects.

In further analyzing the auditory impaired subgroups, a fair alpha coefficient of .59 was determined for males and a fair coefficient of .72 for the older subgroup. A moderate coefficient of .81 was reported for both the younger subgroup and female subgroup. It appears from the results that the stork test is only fairly reliable for the auditory impaired.

These results support the studies of Myklebust (1964), Long (1932/1972), and Morsh (cited in Winnick 1979), that found the deaf inferior to the hearing on static and dynamic balance. This inferior performance could have made the deaf subjects score more inconsistently than the other subjects and may have caused the lower reliability by altering the relative order of performance.

A fair coefficient of .61 was reported for the older orthopedically impaired and a moderate coefficient of .76 for the younger subgroup. However, a high coefficient of .90 was determined for orthopedic females. Males were not evaluated due to insufficient numbers. It should be noted that the number of subjects was low for both the older subgroup and female subgroup (13 & 12). The reported results for this population does not seem unreasonable. Due to the variety of conditions that would decrease balance (asymmetry of body parts, missing limbs, and limb dysfunction) with this group, it seems reasonable that orthopedically impaired subjects would do poorly on tests of balance.

Sit And Reach

Very high alpha coefficients (between .98 & .99) were computed for the sit and reach test on all four groups. These coefficients are better than those reported by Colgan (1978), in the only other study on sit and reach (.84 & .95). Although Colgan's study was only on nonimpaired subjects, the test item still appears to be very reliable for all impaired groups, as well as the nonimpaired.

The Project UNIQUE Manual recommends two trials for the sit and reach test with the mean of both trials being used as the criterion score. According to the results obtained, this may not be the best procedure. The results showed trials one and two to be significantly different for all four groups. In addition, trial two proved to be the better score for all groups (28.5 vs. 29.9, 22.3 vs. 24.2, 22.5 vs. 23.4, and 20.3 vs. 21.5). Therefore, if two trials are continued only trial two should be recorded. Mathews and Fox (1976) and deVries (1974) indicate that a warmup or preliminary stretching increases flexibility. Thus, in order to attain the most accurate measurement, additional trials may be needed or specific stretching exercises should be performed before testing.

Grip Strength

The reliability of the grip strength test item proved to be very high as alpha coefficients between .95 and .99 were reported for all the groups tested. These coefficients compare very well to previous studies done on grip strength. The coefficients in this study are better than those reported by Fleishman (1964b) and Rarick, et. al. (1976) and superior to those reported by Avent (1963) and Keogh (1965). Therefore, the results of this study indicate the grip strength test to be highly reliable.

Analysis of the criterion scores recommended for the grip strength test item points out the possible need for a change in the Project UNIQUE procedure of giving three trials and using the mean of all three trials as the representative score. Only in three cases (right and left hand for visually impaired and right hand for orthopedically impaired) was the mean of

all three trials the appropriate criterion score. In three cases the mean of trials one and two was recommended as the most appropriate score and in two cases trial one was the recommended score (recommended criterion scores for grip strength are listed in Table 5). In addition, an analysis of the trial means showed a decreasing score for all subjects. Therefore, for practical purposes, it might be better to give only one trial and use it as the criterion score.

Flexed Arm Hang

The Project UNIQUE Training Manual recommends giving two trials for the flexed arm hang and using the mean of both trials as the representative score. The results of this study confirm this scoring procedure. Nevertheless, it should be noted that it may be more practical to give only one trial as Colgan (1978) reported high coefficients of .96 and .89 giving one trial on two different days.

The flexed arm hang appears to be a reliable test item, as alpha coefficients of .93--.96 were computed for the non-impaired, auditory impaired, and orthopedically impaired. These coefficients, compare well against prior studies, which reported reliability coefficients between .87--.97 (Vodola, 1978; Bolonchuk, 1971; Colgan, 1978; and Avent 1963). A moderate coefficient of .84 was reported for the visually impaired. A further analysis of the visually impaired revealed alpha coefficients of .80 for the older subgroup and .81 for the male subgroup, while strong coefficients of .90 and .88 were determined for the younger subgroup and female subgroup.

This additional information indicates that although the flexed arm hang is only fairly reliable for older visually impaired students and visually impaired males, it appears reliable for the visually impaired subgroups of younger subjects and females.

Standing Broad Jump

Very high alpha coefficients (.96--.99) were determined for the standing broad jump, indicating it to be a very reliable test item. The obtained coefficients compare very favorably to alpha coefficients reported in several prior studies (see Appendix A). In addition, the alpha coefficients for this study were substantially higher than those reported in the studies by Marmis, et.al. (1969); Keogh (1965); Bolonchuk (1971); Avent (1963) and Colgan (1978).

The results of this study appear to confirm the Project UNIQUE procedure of giving three trials and using the mean of all three trials as the criterion score. However, the results also indicate that using the mean of trials two and three would be a more appropriate criterion measure for the visually impaired. Trial one might possibly be needed as a practice trial for orientating themselves to the testing procedure.

Softball Throw (Time)

Alpha coefficients for the softball throw for time ranged from moderate for the nonimpaired (.84), to moderately high for the visually and auditory impaired (.88), and high for orthopedically impaired (.94). In a study to determine velocity, Rarick, et.al. (1976) obtained higher coefficients (.91--.96), except for a low of .81 on males, ages 6--9. Four trials were given by Rarick et. al. (1976) and this may account for the slightly

higher coefficients. Analysis of the criterion scores confirms the Project UNIQUE procedure of using three trials with the mean of all trials as the criterion score.

Softball Throw (Distance)

The alpha coefficients obtained on the softball throw for distance ranged from a moderately high .87 for orthopedically impaired, a high .95 for nonimpaired, and a very high .99 for visually and auditory impaired. Further analysis revealed fair reliability coefficients of .81 for orthopedically impaired females and .82 for older orthopedically impaired subjects. The younger and male subgroups had high coefficients of .93 and .98. Therefore, it appears that the low reliability might be caused by older orthopedically impaired females. However, this study can not confirm that conclusion.

The alpha coefficients that were determined by this study were equal to the reliability coefficients reported by Marmis, et.al. (1969) and Keogh (1965). As a whole, the alpha coefficients were slightly better than the reliability coefficients reported by Fleishman (1964b), Klesius (1968), and Bolonchuk (1971). Thus, it can be stated that the softball throw for distance is reliable using the Project UNIQUE procedure of giving three trials and using the mean of all three trials as the criterion measure. An analysis of the criterion scores confirms this procedure.

Summary

Based on the results of this study it can be stated that

most of the items proved to be reliable. Only two alpha coefficients were below .70 (rise to stand for nonimpaired and stork test for orthopedically impaired). In addition, there were three coefficients of .70, .76, and .77. All other alpha coefficients were above .80 indicating good to high reliability. Furthermore, the reliability of the test items for the impaired groups was in general equal to or better than the reliability coefficients for the nonimpaired subjects. It should also be noted that the coefficients for this study were, in the majority, superior to those reported in prior studies.

It was surprising that the reliability coefficients for the impaired samples were equal to or better than the nonimpaired group as impaired populations have been found inferior to nonimpaired populations on tests of physical and motor fitness (Sherrill, 1977; Fait, 1978; and Winnick, 1979). Therefore, it was expected that the impaired groups would have lower reliabilities than the nonimpaired subjects. However, the results of this study did not substantiate this belief. This indicated that if the Project UNIQUE testing procedures are followed and the appropriate modifications are made, then the results will be reliable. Therefore, when testing impaired populations on the Project UNIQUE test items, reliable measurements can be expected.

In analyzing the criterion scores to determine the correct scoring procedures, it was discovered that the Project UNIQUE procedures were, in general, appropriate for skinfold measurements, rise to stand, mat creep, shuttle run, stork test, arm

hang, broad jump, and softball throw. However, it was recommended that the Project UNIQUE procedures be changed for the leg raise, trunk raise (nonimpaired), sit and reach, and grip strength. Thus, it may be stated that in most cases the Project UNIQUE procedures were determined to be accurate, however, further research is needed in several areas to confirm this study's conclusions.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary

The major purpose of this study was to determine reliability coefficients and the standard error of measurement for selected health and performance related test items from the Project UNIQUE Physical Fitness Test Inventory. The appropriate number of trials and the correct criterion score were also determined by an analysis of variance and an analysis of the superdiagonal of the inter-trial correlation matrix.

For the multi-trial items, 50 nonimpaired youth, 50 visually impaired youth, 50 auditory impaired youth, and 50 orthopedically impaired youth, between the ages of 10 and 17, were randomly selected from various schools participating in the Project UNIQUE study. Subjects were tested for within day reliability and given three trials on the following items: skinfold measurements, rise to stand, mat creep, modified stork test, grip strength, standing broad jump, and softball throw. For the shuttle run, sit and reach, and flexed arm hang two trials were administered.

Subjects for the single-trial items included 50 nonimpaired youth, 47 visually impaired youth, and 50 auditory impaired youth, between the ages of 10 and 17. These subjects were randomly selected from at least two different sites for each population tested. Sites were located in the Rochester, N.Y.

and Buffalo, N.Y. areas. For both the timed leg raise and timed trunk raise, subjects were tested for among day reliability and given three trials on three consecutive physical education periods, for each test.

Reliability coefficients were determined by using the SPSS subprogram, "Reliability", which resulted in the calculation of Cronbach's alpha coefficient. In determining a criterion score an analysis of variance was calculated with a significant F indicating trend. When a trend was shown among trials, a post hoc analysis was performed to determine which trials were significantly different. The Newman-Keuls technique of post hoc analysis was used and supported by an inspection of the superdiagonal of the correlation matrix.

The results of this study showed the timed leg raise and timed trunk raise to be reliable tests. However, the trunk raise was only moderately reliable for the nonimpaired and needed a much different scoring procedure for the same group. Skinfold measurements were found to be highly reliable for all four groups with no major changes in scoring procedures recommended.

In the area of agility, the results indicated that the rise to stand was reliable and the mat creep highly reliable for all of the impaired groups. However, for the nonimpaired subjects, poor reliability was reported on the rise to stand and moderate reliability on the mat creep. On the shuttle run test, very high reliability was determined for the orthopedically impaired, high reliability for the visually impaired,

moderate reliability for the nonimpaired, and low reliability for the auditory impaired. Slight changes in scoring procedures were recommended for the nonimpaired students on mat creep and shuttle run.

A high coefficient of reliability was determined for non-impaired subjects on the stork test. The results indicated the item to be fairly reliable for the visually impaired and the auditory impaired, but unreliable for the orthopedically impaired. The results also confirmed the procedure of using the mean of all three trials as the criterion score.

This study found the sit and reach test to be a very reliable item. However, it was recommended that trial two was a more appropriate criterion score than the mean of trials one and two. Although it was found to be a highly reliable test, some changes were recommended on the scoring procedure for the grip strength test item. It was suggested that trial one might be the most practical criterion score.

The nonimpaired, auditory impaired, and orthopedically impaired showed good reliability on the flexed arm hang, while the test item proved to be only moderately reliable for the visually impaired. In addition, this study confirmed the procedure of giving two trials with the mean of both trials serving as the criterion score.

The broad jump proved to be highly reliable for all four groups. In addition, only one change was recommended in the Project UNIQUE scoring procedures. For visually impaired subjects the mean of trials two and three represented a better

criterion measure.

Nonimpaired subjects, visually impaired subjects, and auditory impaired subjects showed moderate to good reliability on the softball throw for time test, while the orthopedically impaired showed high reliability. On the softball throw for distance, high to very high reliability coefficients were determined for the nonimpaired, visually impaired, and auditory impaired groups. Good reliability was reported for the orthopedically impaired. No changes were recommended in the Project UNIQUE scoring procedures.

Conclusions

Based on the results of this study the following conclusions were made:

1. Using trend-free trials, the following items are reliable for all populations: skinfold measurements, sit and reach, grip strength, broad jump, and softball throw for distance.

2. Using trend-free trials, the leg raise and trunk raise items are reliable for nonimpaired, visually impaired, and auditory impaired populations.

3. Using trend-free trials the following items are reliable for visually, auditory, and orthopedically impaired populations: rise to stand, mat creep, and softball throw for time.

4. Using trend-free trials the stork stand is reliable for nonimpaired populations, fairly reliable for visually

impaired and auditory impaired but unreliable for orthopedically impaired subjects.

5. Reliability coefficients determined by using trend-free trials were, in the majority, superior to those reported in prior studies.

Recommendations

Based on the results of this study the following recommendations are made:

1. Similar studies are needed with greater classification in regard to sex, age, and impairment.

2. Further research is needed to determine correct procedures for the following test items: leg raise, trunk raise (nonimpaired), sit and reach, and grip strength.

3. Additional research is needed on the reliability of the following items: rise to stand (nonimpaired), mat creep (nonimpaired), shuttle run (auditory impaired), stork stand (all three impaired groups), arm hang (visually impaired), and softball throw for time (nonimpaired).

4. Similar reliability studies are needed to substantiate the results of this study.

"References"

American Alliance for Health, Physical Education, Recreation and Dance. Lifetime health related physical fitness test manual. Reston, Va.: Author, 1980.

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: Author, 1974.

Anhalt, C. J. A reliable and valid motor fitness test battery for upper elementary school girls (Masters thesis, University of Oregon, 1958).

Arnheim, D. D., Auxter, D., & Crowe, W. C. Principles and methods of adapted physical education and recreation. Saint Louis: C. V. Mosbey, 1977.

Avent, H. H. Tests of static and dynamic strength for girls nine through twelve years old (Doctoral dissertation, Indiana University, 1963). University Microfilms, 1963.

Barrow, H. M., & McGee, R. A practical approach to measurement in physical education (3rd ed.). Philadelphia: Lea and Febiger, 1979.

Baumgartner, T. A. Criterion score for multiple trial measures. Reasearch Quarterly, 1974, 45, 193-198.

Baumgartner, T. A. Estimating reliability when all test trials are administered on the same day. Research Quarterly, 1969, 40, 222-225.

Baumgartner, T. A. The applicability of the Spearman-Brown prophecy formula when applied to physical performance test. Research Quarterly, 1968, 39, 847-856.

Baumgartner, T. A., & Jackson, A. S. Measurement for evaluation in physical education. Boston: Houghton Mifflin, 1975.

Baumgartner, T. A., & Jackson, A. S. Measurement schedules for tests of motor performance. Research Quarterly, 1970, 41, 10-14.

Bolonchuk, W. W. A critical analysis of the AAHPER youth fitness test. 1971. (ERIC Document Reproduction Service No. ED 103 414)

Brozek, J., & Alexander, H. Components of variation and the consistency of repeated measurements. Research Quarterly, 1947, 18, 152-166.

Clarke, H. H., & Clarke, D. H. Developmental and adapted physical education (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Colgan, Sharon M. A comparative study of the American Alliance of Health, Physical Education and Recreation youth fitness test and a proposed fitness test (Masters thesis, University of Missouri at Columbia, 1978). Eugene, Or.: University of Oregon, School of HPER, 1980.

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

Dayton, C. M. The design of educational experiments. New York: McGraw Hill, 1970.

deVries, H. A. Physiology of exercise for physical education and athletics. Dubuque, Iowa: Wm. C. Brown, 1974.

Disch, J. Consideration for establishing a reliable and valid criterion measure for a multiple trial motor performance test. In T. A. Baumgartner, J. Safrit, & D. Casady (Eds.), Prodeedings of the C.I.C. symposium on measurement and evaluation in physical education. Bloomington, Indiana: Indiana University, Oct., 1975.

Fait, H. F. Special physical education: Adapted, corrective, developmental (4th ed.). Philadelphia: W. B. Saunders, 1978.

Final rules and regulations of the education of handicapped children, implementation of part B of the education of the handicapped act. Federal Register, August 23, 1977, 42, No. 163, Part II.

Fleishman, E. A. Examiner's manual for the basic fitness tests. Englewood Cliffs, N.J.: Prentice-Hall, 1964a.

Fleishman, E. A. The structure and measurement of physical fitness. Englewood Cliffs, N.J.: Prentice-Hall, 1964b.

Franks, B. D., & Deutsch, H. Evaluating performance in physical education. New York: Academic Press, 1973.

Jackson, A., Jackson, A. S., & Bell, J. A comparison of alpha and the intraclass reliability coefficients. Research Quarterly, 1980, 51, 568-571.

Jackson, A., Polock, M. L., & Gettman, L. R. Intertester reliability of selected skinfold and circumference measurements and percent fat estimates. Research Quarterly, 1978, 49, 546-551.

Kane, R. J., & Meredith, H. V. Ability in the standing broad jump of elementary school children 7, 9, and 11 years of

age. Research Quarterly, 1952, 23, 198-208.

Kelly, T. L. Interpretation of Educational Measurements. Yonkers-on-Hudson, N.Y.: World Book Company, 1927.

Keogh, J. Motor performance of elementary school children. Los Angeles: Physical Education Department, University of California at Los Angeles, 1965.

Kirkendal, D. R., Gruber, J. J., & Johnson, R. E. Measurement and evaluation for physical educators. Dubuque, Iowa: Wm. C. Brown, 1980.

Klesius, S. E. A comparison of methods for estimating the reliability of selected test items of the AAHPER youth fitness test (Masters thesis, Florida State University, 1966). Eugene, Or.: University of Oregon, School of HPER, 1966.

Klesius, S. E. Reliability of the AAHPER youth fitness test items and relative efficiency of the performance measures. Research Quarterly, 1968, 39, 809-811.

Kraus, H., & Hirschland, R. P. Minimum muscular fitness tests in school children. Research Quarterly, 1954, 25, 187-188.

Kroll, W. A note on the coefficient of intraclass correlation as an estimate of reliability. Research Quarterly, 1962, 33, 313-316.

Kroll, W. Reliability theory and research decision in selection of a criterion score. Research Quarterly, 1967, 38, 412-419.

Kroll, W. Test reliability and errors of measurement at several levels of absoulute isometric strengths. Research Quarterly, 1970, 41, 155-163.

Long, J. A. Motor abilities of deaf children. New York: AMS Press, Inc., 1972. (Originally published, 1932.)

Marmis, C., Montoye, H. J., Cunningham, D. A., & Rozar, A. Reliability of the multi-trial items of the AAHPER youth fitness test. Research Quarterly, 1969, 40, 240-245.

Mathews, D. K., & Fox, E. L. The physiological basis of physical education and athletics. Philadelphia: W. B. Saunders, 1976.

Myklebust, H. R. The psychology of deafness. New York: Grune & Stratton, Inc., 1964.

Proposed rules and regulations for Public Law 94-142. Federal Register, December 30, 1976, 41, No. 252, 56977.

Rarick, G. L. (Ed.). Physical activity: Human growth and development. New York: Academic Press, 1973.

Rarick, G. L., Dobbins, D. A., & Broadhead, G. D. The motor domain and its correlates in educationally handicapped children. Englewood Cliffs, N.J.: Prentice-Hall, 1976.

Safrit, M. J. Evaluation in physical education (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1981.

Safrit, M. J. (Ed.). Reliability theory. Washington, D.C.: American Alliance for Health, Physical Education and Recreation, 1976.

Scott, M. G. (Ed.). Research methods in health, physical education and recreation (2nd ed.). Washington, D.C.: American Alliance for Health, Physical Education and Recreation, 1959.

Sherrill, C. Adapted physical education and recreation: A multidisciplinary approach. Dubuque, Iowa: Wm. C. Brown, 1977.

Sinning, W. E. Use and misuse of anthropometric estimates of body composition. Journal of Physical Education and

Recreation, Feb. 1980, pp. 43-45.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Thorndike, R. L. Reliability. In E. F. Lindquist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1951.

Vodola, T. M. Developmental and adapted physical education (ACTIVE, motor ability and physical fitness norms for normal, mentally retarded, learning disabled, and emotionally disturbed individuals). Oakhurst, N.J.: Township of Ocean School District, 1978.

Wilmore, J., & Behnke, A. R. An anthropometric estimation of body density and lean body weight in young men. Journal of Applied Physiology, 1969, 27, 25-31.

Wilmore, J., & Behnke, A. R. An anthropometric estimation of body density and lean body weight in young women. American Journal of Clinical Nutrition, 1970, 23, 167-274.

Winnick, J. P. Early movement experiences and development: Habilitation and remediation. Philadelphia: W. B. Saunders, 1979.

Winnick, J. P., & Short, F. X. Project UNIQUE, the physical fitness and performance of sensory and orthopedically impaired youth: A training manual for site coordinators and testers. Brockport, N.Y.: SUNY-College at Brockport, 1980.

APPENDICES

APPENDIX A

SUMMARY TABLE OF PRIOR STUDIES

ITEM	SOURCE	SAMPLE	TYPE/ PROCEDURE	TRIALS (t) DAYS (d)	REL. COEF.
Leg raise	Fleishman (1964a)	201 Navy recruits avg age 18 yrs 3 mths	test-retest/ correl. coef.[1]	not given	.71
Trunk raise	Rarick, et.al. (1976)	71 males ages 6-9	within day test-retest / correl. coef.	2t	.748
		65 retarded males ages 10-13			.822
		71 retarded males 6-9			.779
		74 females ages 6-9			.705
		61 retarded females ages 10-13			.804
		64 retarded females ages 6-9			.843
	Avent(1963)	50 females ages 9-12	test-retest/ Pearson product-moment	1t 2d 3-8 wks apart	.458
Skinfold measurements	Colgan(1978)	164 males & 162 females ages 10-18	between day test-retest/ correlation coefficients	1t 2d 1 wk apart	Triceps: .93f .96m Subscapular .94f .90m
Mat creep	Rarick, et.al. (1976)	71 males ages 6-9	within day test-retest/ correlation coefficients	3t 1d	.925
		65 retarded males ages 10-13			.918
		71 retarded males ages 6-9			.924
		74 females ages 6-9			.940
		61 retarded females ages 10-13			.939
		64 retarded females ages 6-9			.949

[1] correlation coefficients

ITEM	SOURCE	SAMPLE	TYPE/ PROCEDURE	TRIALS(t) DAYS (d)	REL. COEF.
Shuttle run	Fleishman (1964b)	20,000 males & females ages 12-18	test-retest/ correl. coef.	not given	.85
	Klesius (1968)	150 10th grade males	test-retest/ correl. coef.	3t	.68
	Keogh (1965)	24 1st grade males & females 24 3rd grade males & females	test-retest/ Pearson product-moment	2t 2d 2-6 wks apart	.73 .59
	Baumgartner & Jackson (1970)	95 Junior-high boys 82 Senior-high boys	int. con.*/ ANOVA	4t	.76 .82
	Baumgartner (1974)	95 Junior-high boys 82 Senior-high boys	int. con./ ANOVA	4t 2d	$.81^1$ $.93^2$ $.87^1$ $.92^2$
	Marmis,et.al. (1969)	1,122 males ages 9-18 938 females ages 9-18	test-retest/ correl. coef.	2t	between: .60-.80 between: .46-.82
	Colgan (1978)	164 males ages 10-18 162 females ages 10-18	test-retest/ correl. coef.	2t 2d 1 wk apart	.82 .90
	Anhalt (1958)	32 4th-5th-6th grade females	between day test-retest/ Pearson product-moment	2t 2d 1 wk apart	.887

[1]sum of scores used

[2]best score used

*internal consistency

ITEM	SOURCE	SAMPLE	TYPE/ PROCEDURE	TRIALS (t) DAYS (d)	REL. COEF.
Shuttle run (cont'd)	Rarick, et.al. (1976)	71 males ages 6-9	within day test-retest / correlation coefficients	not given	.955
		65 retarded males ages 10-13			.915
		71 retarded males ages 6-9			.940
		74 females ages 6-9			.947
		61 retarded females ages 10-13			.905
		64 retarded females ages 6-9			.926
Stork test	Fleishman (1964b)	20,000 males & females ages 12-18	test-retest/ correl. coef.	not given	.82
	Rarick,et.al. (1976)	71 males ages 6-9	within day test-retest/ correlation coefficients	3t	.790
		65 retarded males ages 10-13			.781
		71 retarded males ages 6-9			.791
		74 females ages 6-9			.798
		61 retarded females ages 10-13			.727
		64 retarded females ages 6-9			.818
Sit and Reach	Colgan (1978)	164 males ages 10-18	between day test-retest/ correl. coef.	1t 2d 1 wk apart	.84
		162 females ages 10-18			.95

ITEM	SOURCE	SAMPLE	TYPE/ PROCEDURE	TRIALS (t) DAYS (d)	REL. COEF.
Grip Strength	Fleishman (1964b)	20,000 males & females ages 12-18	test-retest/ correl. coef.	not given	.91
					rt. lft.
	Rarick,et.al. (1976)	71 males ages 6-9	within day test-retest/ correlation coefficients	3t	.911 .959
		65 retarded males ages 10-13			.927 .941
		71 retarded males ages 6-9			.902 .917
		74 females ages 6-9			.882 .896
		61 retarded females ages 10-13			.975 .959
		64 retarded females ages 6-9			.917 .934
	Keogh (1965)	23 1st grade males & females	test-retest/ Pearson product moment	2t 2d 2-6 wks apart	.85 .79
		23 3rd grade males & females			.75 .70
					right hand only
	Keogh (1965)	23 1st grade males & females	within day test-retest/Pearson product-moment	2t	.76
		23 3rd grade males & females			.84
	Avent (1963)	50 females ages 9-12	between day test-retest/Pearson product-moment	2t 2d 3-8 wks apart	.654 .797
Flexed Arm hang	Vodola (1978)	30 females age 15	between day test-retest/Pearson product-moment	one day between tests	.97
		33 males age 15			.89
		19 females age 7			.88
	Bolonchuk (1971)	25 5th & 6th grade females	test-retest/ Pearson product-moment	2t	.95
	Colgan (1978)	164 males & 162 females ages 10-18	between day test-retest/ correl. coef.	1t 2d 1 wk apart	.96m .89f

ITEM	SOURCE	SAMPLE	TYPE/ PROCEDURE	TRIALS (t) DAYS (d)	REL. COEF.
Flexed Arm hang (cont'd)	Avent (1963)	50 females ages 9-12	between day test-retest/Pearson product-moment	2t 2d 3-8 wks apart (underhand grip used)	.868
Standing Broad jump	Klesius (1968)	150 10th grade males	test-retest/ correl. coef.	3t	.94
	Marmis,et.al. (1969)	1,122 males ages 9-18 938 females ages 9-18	test-retest/ correl. coef.	3t	.73-.95 .75-.95
	Vodola (1978)	30 females age 15 33 males age 15 13 males age 6 19 females age 7	between day test-retest/Pearson product-moment	one day between tests	.95 .98 .49 .89
	Rarick,et.al. (1976)	71 males ages 6-9 65 retarded males ages 10-13 71 retarded males ages 6-9 74 females ages 6-9 61 retarded females ages 10-13 64 retarded females ages 6-9	within day test-retest/ correlation coefficients	4t	.805 .917 .947 .906 .953 .957
	Kane & Meredith (1952)	300 males ages 7,9,11 300 females ages 7,9,11	within day test-retest/Pearson product-moment	12t (best trial & second best trial)	.97-.99 .98-.99
	Kane & Meredith (1952)	75 males age 7 75 females age 7	between day test-retest/Pearson product-moment	12t 2d (best score on both days)	.83 .86

ITEM	SOURCE	SAMPLE	TYPE/ PROCEDURE	TRIALS (t) DAYS (d)	REL. COEF.
Standing Broad jump (cont'd)	Keogh (1965)	21 1st grade males & females 27 3rd grade males & females	between day test-retest/Pearson product-moment	3t 2d 2-6 wks apart	.90 .77
	Keogh (1965)	Not given	within day test-retest/Pearson product-moment	3t (best & 2nd best used)	.91
	Baumgartner & Jackson (1970)	95 Junior-high males 82 Senior-high males	int. con./ ANOVA	6t	.96 .97
	Baumgartner (1974)	95 Junior-high males 82 Senior-high males	int. con./ ANOVA	6t 2d	.96[1] .96[2] .95[1] .96[2]
	Bolonchuk (1971)	20 5th & 6th grade males 25 5th & 6th grade females	test-retest/ Pearson product-moment	3t	.89 .82
	Avent (1963)	50 females ages 9-12	between day test retest/Pearson product-moment	3t 2d 3-8 wks apart	.681
	Anhalt (1958)	32 4th-5th-6th grade females	between day test-retest/Pearson product-moment	3t 2d 1 wk apart	.913
	Colgan (1978)	164 males ages 10-18 162 females ages 10-18	between day test-retest/ correl. coef.	3t 2d 1 wk apart	.81 .82

[1] sum of scores used

[2] test score used

ITEM	SOURCE	SAMPLE	TYPE/ PROCEDURE	TRIALS (t) DAYS (d)	REL. COEF.
Softball throw (distance)	Fleishman (1964b)	20,000 males & females ages 12-18	test-retest/ correl. coef.	not given	.93
	Klesius (1968)	150 10th-grade males	test-rest/ correl. coef.	3t	.93
	Marmis,et.al. (1969)	1,122 males ages 9-18 938 females ages 9-18	test-retest/ correl. coef.	3t	.86-.94 .83-.97
	Keogh (1965)	19 1st-grade males & females 27 3rd-grade males & females	between day test-retest/Pearson product-moment	3t 2d 2-6 wks apart	.97 .88
	Keogh (1965)	Not given	within day test-retest/Pearson product-moment	3t (best & 2nd best trial used)	.95
	Bolonchuk (1971)	20 5th & 6th grade males 25 5th & 6th grade females	test-retest/ Pearson product-moment	3t	.94 .93
Softball throw (timed)	Rarick,et.al. (1976)	71 males ages 6-9 65 retarded males ages 10-13 71 retarded males ages 6-9 74 females ages 6-9 61 retarded females ages 10-13 64 retarded females ages 6-9	within day test-retest/ correlation coefficients	4t	.863 .964 .966 .709 .950 .854

APPENDIX B

COMPARISON OF THE INTRACLASS COEFFICIENT AND ALPHA COEFFICIENT

ANOVA SUMMARY TABLE

Source of Variation	SS	DF	MS	F
Between people	14020.48835	49	286.13242	
Within people	84.96000	100	0.84960	
Between measures	2.22013	2	1.11007	1.31480
Residual	82.73987	98	0.84428	
Total	14105.44835	149	94.66744	

COVARIANCE MATRIX

Trial	1	2	3
1	96.66211	96.38975	94.40149
2	96.38975	96.85708	94.49689
3	94.40149	94.49689	94.30180

INTRACLASS

$$\underline{R} = \frac{MS_s - MS_{sxt}}{MS_s}$$

$$= \frac{286.13242 - .84428}{286.13242}$$

$$= .9970493$$

ALPHA

$$\alpha = \frac{k}{k-1}\left(1 - \frac{S_t^2}{S_x^2}\right)$$

$$= \left(\frac{3}{2}\right)\left(1 - \frac{287.82099^{1}}{858.39725^{2}}\right)$$

$$= \left(\frac{3}{2}\right)\left(1 - .3353005\right)$$

$$= 1.5 \times .6646995$$

$$= .9970493$$

[1] sum of the main diagonal

[2] total test variance (sum of all 9 values)

APPENDIX C

SUMMARY DATA CALCULATED USING ALL TRIALS

Variable	Trials	Nonimpaired			Visual			Auditory			[illegible]		
		Grand M	Rel. Coef.	SE_m	Grand M	Rel. Coef.	SE_m	Grand M	Rel. Coef.	SE_m	Grand M	Rel. Coef.	SE_m
L Raise*	3	40.52	.83	9.53	41.99 (47[1])	.91	11.34	53.93	.93	9.12	---	---	---
T Raise*	3	61.35	.91	13.33	47.75 (46[1])	.99	4.57	59.37	.91	11.40	---	---	---
Triceps	3	12.72	.99	.33	13.75	.97	1.04	12.49	.98	.78	12.70	.97	.85
Abdom	3	14.14	.99	.54	16.68	.90	2.66	13.74	.98	.91	13.93 (49[1])	.99	.53
Subscp	3	10.29	.99	.36	13.84	.99	.60	10.45	.99	.44	10.64	.99	.35
Stand	3	1.38	.40	.38	1.88	.91	.22	1.67	.90	.16	4.39 (25[1])	.93	.51
M Creep	3	3.56	.85	.15	4.42	.97	.23	3.86	.96	.20	10.23 (41[1])	.97	1.22
SR Time	2	11.31	.86	.46	12.49	.92	.68	11.95	.70	2.48	34.77	.99	2.04
Stork	3	46.96	.89	26.92	17.15	.77	10.06	9.89	.76	7.52	2.42 (22[1])	.61	2.86
S Reach	2	29.19	.99	.95	23.22	.98	1.15	22.95	.99	.94	20.93 (38[1])	.99	.97
R Grip	3	24.01	.98	1.08	23.58	.99	1.14	23.42	.97	1.36	14.15	.98	1.33
L Grip	3	23.37	.98	1.04	21.17	.97	1.69	21.01	.98	1.21	14.86	.99	1.00
Arm Hang	2	8.35	.93	2.88	11.48	.84	4.85	9.17	.92	3.11	2.50	.96	1.08
B Jump	3	5.27	.96	.18	4.96	.94	.29	5.08	.98	.16	2.04 (21[1])	.99	.16
SB Time	3	2.13	.84	.24	1.88	.88	.26	2.03	.88	.22	1.05 (49[1])	.94	.14
SB Dist	3	92.39	.95	7.62	70.71	.99	4.35	84.73	.99	4.11	28.91	.86	4.51

*Single-trial between days

[1]number of subjects

APPENDIX D

SUMMARY DATA OF NONIMPAIRED SUBGROUPS

Variable	Trials	Young Grand M	∝ Coef.	N	Grand M	∝ Coef.	N	Grand M	∝ Coef.	N	Grand M	∝ Coef.	N
L Raise*	3	41.27+	.83	47	insufficient numbers			47.65+	.81	20	35.77	.84	30
T Raise*	3	58.96	.89	31	65.26	.93	19	72.02	.93	20	54.24	.88	30
Triceps	3	12.45	.99	37	13.48	.99	13	11.52	.99	14	13.18	.99	36
Abdom	3	14.05	.99	37	14.40+	.99	13	11.11	.99	14	15.32	.99	36
Subscp	3	10.72	.99	37	9.05	.99	13	9.46	.99	14	10.61	.99	36
Stand	3	1.40	.37	37	1.3	.88	13	1.39	.79	14	1.38	.37	36
M Creep	3	3.53+	.82	37	3.64	.90	13	3.47	.89	14	3.59+	.84	36
SR Time	2	11.46+	.83	37	10.87	.93	13	11.59	.79	14	11.19+	.89	36
Stork	3	50.74	.89	37	36.21	.67	13	22.95	.73	14	56.30	.89	36
S Reach	2	27.74+	.99	37	33.31+	.98	13	23.57+	.98	14	31.37+	.98	36
R Grip	3	22.33+	.97	37	28.77+	.98	13	19.29	.96	14	25.84+	.97	36
L Grip	3	21.05+	.98	37	26.15	.97	13	17.90	.96	14	24.11+	.98	36
Arm Hang	2	6.97+	.93	37	12.27	.92	13	8.57	.88	14	8.26	.94	36
B Jump	3	5.06	.93	37	5.90	.96	13	4.93	.92	14	5.41+	.96	36
SB Time	3	2.02	.77	37	2.45	.91	13	2.26	.70	14	2.09	.87	36
SB Dist	3	82.96	.93	37	119.23	.94	13	97.77	.94	14	96.30	.95	36

*single trial between days

+significant F was reported

APPENDIX E

SUMMARY DATA OF VISUALLY IMPAIRED SUBGROUPS

Variable	Trials	Young Grand M	Young ∝ Coef.	Young N	Old Grand M	Old ∝ Coef.	Old N	Mixed Grand M	Mixed ∝ Coef.	Mixed N	Total Grand M	Total ∝ Coef.	Total N
L Raise*	3	I N S U F F			I C I E N T			N U M B E R S					
T Raise*	3	I N S U F F			I C I E N T			N U M B E R S					
Triceps	3	15.20	.99	13	13.24	.96	37	12.36	.98	28	15.53	.94	22
Abdom	3	18.28	.89	13	16.11	.90	37	16.28	.85	28	17.25	.99	22
Subscp	3	13.08	.97	13	14.11+	.91	37	13.80	.99	28	13.90	.97	22
Stand	3	1.99	.87	13	1.84	.93	37	2.03	.91	28	1.68	.91	22
M Creep	3	4.76+	.98	13	4.30+	.97	37	4.11	.97	28	4.81	.97	22
SR Time	2	12.80	.94	13	12.38	.90	37	11.82	.91	28	13.33	.89	22
Stork	3	10.15	.86	13	19.61	.74	37	15.02	.78	28	19.86	.75	22
S Reach	2	22.69+	.95	13	23.41+	.99	37	20.00	.98	28	27.32	.98	22
R Grip	3	19.56+	.98	13	24.99	.99	37	26.33	.99	28	20.08	.97	22
L Grip	3	17.92	.94	13	22.31+	.97	37	23.36	.97	28	18.39	.95	22
Arm Hang	2	11.96	.90	13	11.31	.80	37	14.75	.81	28	7.32	.88	22
B Jump	3	4.24	.96	13	5.21	.93	37	5.32+	.92	28	4.50	.98	22
SB Time	3	1.62	.87	13	1.98	.88	37	2.03	.93	28	1.70	.78	22
SB Dist	3	57.40	.97	13	75.40	.99	37	85.36	.99	28	52.07	.97	22

*single-trial between day items

+significant F was reported

APPENDIX F

SUMMARY DATA OF AUDITORY IMPAIRED SUBGROUPS

Variable	Trials	Young Grand M	Young ∝ Coef.	Young N	Old Grand M	Old ∝ Coef.	Old N	Male Grand M	Male ∝ Coef.	Male N	Female Grand M	Female ∝ Coef.	Female N
L Raise*	3	I N S U F			F I C I E N T			N U M B E			R S		
T Raise*	3	I N S U F			F I C I E N T			N U M B E			R S		
Triceps	3	14.21+	.97	28	10.30	.99	22	10.78	.98	24	14.07	.98	26
Abdom	3	15.40+	.99	28	11.62	.97	22	12.20	.99	24	15.16+	.97	26
Subscap	3	10.45	.99	28	10.44+	.99	22	9.06	.99	24	11.73+	.99	26
Stand	3	1.62	.86	28	1.74	.92	22	1.73	.91	24	1.62	.87	26
M Creep	3	4.13	.97	28	3.52	.88	22	3.76	.93	24	3.94	.98	26
SR Time	2	11.85	.89	28	12.07	.70	22	11.21	.80	24	12.63	.69	26
Stork	3	9.87	.81	28	9.92	.72	22	9.64	.59	24	10.13	.81	26
S Reach	2	22.50+	.99	28	23.52+	.99	22	19.81+	.99	24	25.85+	.99	26
R Grip	3	19.29+	.97	28	28.68	.96	22	26.67	.98	24	20.42+	.95	26
L Grip	3	17.08	.95	28	26.02	.98	22	24.11	.98	24	18.14	.96	26
Arm Hang	2	4.80+	.90	28	14.72	.87	22	14.23+	.93	24	4.50	.64	26
B Jump	3	4.44	.94	28	5.89	.98	22	5.60	.99	24	4.59	.95	26
SB Time	3	1.92	.83	28	2.17	.93	22	2.35+	.87	24	1.73	.81	26
SB Dist	3	69.90	.99	28	103.60	.99	22	109.29	.99	24	62.05	.96	26

*single-trial between day items

+a significant F was reported

APPENDIX G

SUMMARY DATA OF ORTHOPEDICALLY IMPAIRED SUBGROUPS

Variable	Trials	Young Grand M	α Coef.	N	Old Grand M	α Coef.	N	Male Grand M	α Coef.	N	Female Grand M	α Coef.	N
L Raise*	3	INSUFFICIENT NUMBERS			INSUFFICIENT NUMBERS								
T Raise*	3	INSUFFICIENT NUMBERS			INSUFFICIENT NUMBERS								
Triceps	3	11.19	.98	26	14.33	.97	24	11.26	.96	26	14.25	.99	24
Abdom	3	12.04	.99	26	16.06	.99	24	13.04	.99	26	14.93	.99	23
Subscp	3	8.85	.99	26	12.58	.99	24	9.48	.99	26	11.90	.99	24
Stand	3	4.45	.92	13	4.33	.96	12	3.98	.94	13	4.85	.92	12
M Creep	3	11.07	.98	21	9.35	.95	20	9.88	.97	21	10.59	.98	20
SR Time	2	37.57	.99	26	31.73	.99	24	30.85	.99	26	39.01	.99	24
Stork	3	2.56	.76	26	2.33	.61	13	insufficient number			1.06	.90	12
S Reach	2	20.05	.99	19	21.82	.99	19	20.16	.99	19	21.71	.99	19
R Grip	3	9.15	.98	26	19.57	.98	24	15.44	.99	26	12.76	.98	24
L Grip	3	9.62+	.98	26	20.54	.99	24	17.04+	.99	26	12.50+	.98	24
Arm Hang	2	1.10	.93	26	4.02	.96	24	3.92	.96	26	.96	.93	24
B Jump	3	insufficient numbers			2.27	.99	14	insufficient numbers			1.23	.99	11
SB Time	3	.87	.81	25	1.23	.97	24	1.20	.85	25	.89	.99	24
SB Dist	3	18.36	.93	26	40.33	.82	24	31.90	.98	26	25.66	.81	24

*single-trial between day items

+a significant F was reported

APPENDIX H

DESCRIPTION OF TESTS AND TEST PROCEDURES

The tests and testing procedures employed in this study are described in this appendix. This description includes a brief explanation of the test items and the scoring procedures. It must be stated that these descriptions are a summarization of the testing procedures in the Project UNIQUE Training Manual (Winnick & Short, 1980) and the reader should refer to the manual if a more extensive description is needed.

If fatigue appeared to be influencing a subject's performance, testers were instructed to provide rest intervals or terminate testing the subject. Testers were also advised to demonstrate that item to subjects. A positive approach toward subjects participating in this study was emphasized. Testers were instructed to encourage subjects to try as hard as possible and to give verbal reinforcement after each trial on an activity. Enthusiastic encouragement by the tester during the administration of test items was emphasized so that each subject's participation in the testing would be a positive experience. On the other hand, testers were encouraged to discourage razzing or cheering by observers (including subjects in the group) in all performances.

Timed Leg Raise

In the leg raise, the subject was tested to determine the length of time that straightened legs could be held from a supporting surface. This test item was begun from a supine position with clasped hands placed behind the head/neck area, the elbows flat on the floor, legs straight, and both feet and legs together. From this position, both legs were elevated to approximately 12 inches above the floor and held for as long as possible. If the subject's legs bent, separated, or became heightened or lowered more than three inches above the 12-inch mark, the timing of this activity ceased. It was recommended that testers place a one-foot ruler on the floor under the subject's heels to determine height. Once timing began, it was recommended that the ruler be moved to the side and be used as a guide. The subject's score was the number of seconds that the subject's legs were held in the desired position. Three trials were given for this activity.

Timed Trunk Raise

In this test, the subject was tested to determine the length of time that the hyperextended trunk could be held in a raised position from the prone. The starting position for this test item was a prone position on a gym mat in such a way that the upper body above the illiac crest (belt line) protruded beyond the edge of the mat. The fingers were

clasped and placed behind the head with the elbows pointed outward and the calves of the subject were held down. From this position, the subject hyperextended the back and attempted to hold this position for as long as possible. Timing ceased when the subject lowered the trunk to approximately four inches from the floor. Testers were encouraged to place a 2" X 2" X 4" wooden block to the side of each subject to serve as a guide for the four-inch criterion for the cessation of timing. The subject's score on this test item was the number of seconds that the subject was able to hold the required hyperextended position. Each subject performed three trials in this test item. A brief practice trial was permitted.

Skinfold Measurements

Skinfold measures were taken at the triceps, subscapular, and abdominal regions. Testers obtained three readings at each site for each subject. A green felt tip pen was used to mark a dot at the exact spot at which measures were taken. Skinfolds were obtained by grasping the skin and underlying fat between the thumb and index finger with the span of the grasp dependent upon the thickness of the skinfold. The skinfold caliper was applied approximately one centimeter (less than one-half inch) above the fingers holding the skinfold. All skinfold measurements were taken in the following order: triceps, subscapular, abdomen. This order was repeated for the second and third measurements. Skinfold

measurements were recorded to the nearest millimeter. Three readings were obtained at each site for each subject. In addition, the mean of the three site readings were recorded.

Triceps. The triceps skinfold was taken at the back of the dominant arm midway between the elbow and the apex of the armpit. With the subject's arm freely hanging, the skinfold was taken parallel to the long axis of the arm. The triceps skinfold was a vertical fold.

Subscapular. The subscapular skinfold was taken at a site one inch below and medial to the inferior angle of the scapula on the dominant side. The subscapular skinfold was taken at an angle (in line with the natural cleavage lines of the skin).

Abdominal. The abdominal skinfold was a vertical fold taken at a site two inches to the right of the person's midline in line with the umbilicus, and parallel to the long axis of the body.

Rise to Stand

In this test item, subjects were asked to move from a supine position on a mat to a stable standing position as quickly as possible. The supine position on a mat was the starting position for the rise-to-stand. The hands of each subject were placed by their side, palms down. From this position, the subject moved to a standing position as quickly as possible. To signal, testers raised their hand above their head in clear view to the subject and said READY.

When the hand was dropped to the thigh area and the instructor said GO, the subject began. Testers ceased timing when the subject came to a stable standing position. Each subject was administered three trials. The time elapsed (correct to the nearest tenths of a second) in assuming a standing position from a supine position was recorded. The mean of the three trial scores was also recorded.

Mat Creep

In this test, subjects creeped on their hands and knees on a mat as quickly as possible from a starting line to and around a pylon eight feet away and then back to the finish line. Subjects were required to creep rather than pivot around the pylon. Knee pads could be worn by subjects to prevent abrasions. The subjects were signaled to begin by having the tester raise their hand above their head, verbally commanding READY, and dropping their hand to their thigh, verbally commanding GO! The stopwatch was started on the GO command and stopped as soon as any body part broke the plane of the finish line on the mat. Subjects were provided three trials. The subject's trial score was the time (correct to the nearest tenth of a second) that it took a subject to complete the mat creep. The mean of trial scores was recorded.

Shuttle Run

In the shuttle run subjects ran 30 feet from a starting line, picked up a 2" X 2" X 4" block, ran back to the starting

line, placed the block behind the starting line, ran 30 feet, picked up another block, and ran back to the starting line (subjects were not required to place the second block on the ground or floor). Thus, the start and finish lines were one in the same. Testers were instructed to test two subjects running in different lanes at one time. Each runner was tested in a lane approximately 15 feet in width and 30 feet in length (plus end line space). All subjects were required to wear athletic sneakers and be tested on a hard flat surface. Testers were instructed to use both verbal and hand signals to signal subjects. Subjects falling during the test were retested after a brief interval. In case of a false start, subjects were retested immediately. Subjects were permitted two trials for this test item with at least a three-minute rest interval between trials. The trial score for the shuttle run was the amount of time, correct to tenths of a second, that it took to run the complete shuttle course. The mean of trial scores was also recorded.

Modified Stork Test

In the modified stork test, the subject attempted to balance in a standing position with the arms folded, eyes closed, and one leg raised off the supporting surface by bending the knee. Testers were instructed to place the subject away from obstacles or pointed objects, provide a demonstration of correct procedures, and test at least two subjects at a time. Minute pivoting on the standing leg was

permitted to the extent the tester felt that no significant advantage was provided and the individual remained in essentially the same spot. Stopping or placing the raised leg in contact with the standing leg was not permitted. Timing on each trial ceased if the subject allowed the elevated foot to touch the floor, the subject opened eyes, or if their arms unfolded. Testers were instructed to demonstrate the modified stork test to the subject, allow the subjects a brief practice trial, and demonstrate acceptable and unacceptable pivoting.

Each subject was administered three trials. The trial score on the modified stork test was the balancing time, correct to the nearest second, that the individual was able to hold the desired position. The time recorded for the modified stork test was correct to the nearest second. The mean of the three trials was also recorded.

Sit and Reach Test

In the sit and reach test, each subject was asked to reach forward as far as possible from a sitting position. To start the test, each subject was instructed to lie supine with legs straight, feet placed flush against the sandpapered side of a sit and reach apparatus (AAHPERD, 1980) with shoes removed. From this position, the subject moved to a sitting position. A partner then held down the subject's knees firmly so that they remained in contact with the floor or supporting surface. The subject then extended the arms and hands in

front of the body (one hand on top of the other) and attempted to reach past their toes and contact the centimeter measurement board on the sit and reach apparatus with their fingertips. Subjects were cautioned not to bob with their torso, but rather to gradually reach as far forward as possible and hold for one second. Testers were instructed to provide a practice trial for each subject. Two trials of this test item were administered. The trial score on the sit and reach test was the distance, to the nearest centimeter, that the subject was able to reach and hold for one second. Trial scores, as well as the mean of two trial scores, were recorded for each subject.

Grip Strength

In grip strength tests, subjects squeezed a hand dynamometer with maximum force. Right and left grip strength was measured with the use of an adjustable hand grip dynamometer (Preston PC5032). The dynamometer was adjusted for each subject so that the middle joint (second joint) of the fingers fit firmly around the pulling mechanism as the heel of the hand was placed at the base of the dynamometer and the thumb was wrapped around the base. The subjects squeezed the dynamometer while seated in a straight back desk chair. Hand and arm contact with the seat or any other obstacle which might provide additional leverage or impede movement was not permitted. At the signal to begin, the subject squeezed the dynamometer as hard as possible with the arm

extended downward.

The score for each trial was recorded to the nearest kilogram. The average of the three trials for the right hand and the average of the three trials with the left hand were recorded. Three trials were given for each subject with each hand. Right and left hand trials were alternated as subjects were tested.

Flexed Arm Hang

In the flexed arm hang, the subject grasped a horizontal bar using an overhand (pronated) grip, the elbows flexed, the chest close to the bar, and the chin over the bar, and attempted to hold this position for as long as possible. If the chin of the subject rested on the bar, the subject tilted the head back in an attempt to keep above the bar, or the subject's chin fell below the bar, timing of this activity ceased. Each subject was provided two trials on the flexed arm hang with a minimum of one minute rest interval provided between the trials. Testers were encouraged to administer the second trial after a complete class was tested to allow rest between trials. It was recommended the spotters be used to help subjects assume the proper starting position and to prevent injury from falling. However, spotters were not permitted to touch subjects as they performed the test item. Two trials were administered to each subject. The time, correct to the nearest second, that the subject correctly maintained the flexed arm hang position was the trial score

for each subject. The subject's score on this test item was the mean score of the two trials.

Standing Broad Jump

In the standing broad jump, subjects were directed to jump as far as possible from a take-off line using a two foot take-off and landing on both feet. The jump was measured from the take-off line to the point nearest the take-off line where the subject's back heel touched the floor or ground. It was recommended that each subject be permitted a practice trial during which measuring procedures and the importance of the arm swinging and body rocking for maximum performance were explained. Scoring for the standing broad jump consisted of recording the distance the subject jumped (heel mark) in feet and inches to the nearest rounded inch. Each subject was provided three trials. The subject's score in this test item was determined by computing the mean score of the three trials.

Softball Throw

For this test item, each subject was instructed to throw a regulation softball overhand at an angle of approximately 40 degrees as far forward as possible. Subjects were permitted one forward step during the overhand throwing motion. A stopwatch was started at the time that the ball was released from the subject's hand and stopped when the ball landed on the ground. It was recommended that all subjects be permitted

to warm-up before throwing any distances (practice or actual). Testers were asked not to test subjects during excessive wind conditions (15 mph or more).

Each subject was allowed two practice throws and three test throws. Two scores were recorded for each softball throw trial. The first score was a timed score in tenths of a second and represented the flight time of the subject's throw. The second score was a distance score. It represented the actual feet and inches that the ball traveled. Averaged trial times and trial distance scores were also recorded on this test item.

TEST MODIFICATIONS

Since both impaired and nonimpaired populations were tested in this study, modifications in some of the test items were made in order to have maximum participation. While the modifications are guidelines, they were developed in consultation with physical activity experts in each of the impairment areas (Winnick & Short, 1980). Close conformity with the modifications facilitated safe testing and standardization of procedures. Once again, it must be noted that this is a brief summary of the test modifications that appear in the Project UNIQUE Training Manual. If further detail is needed the reader should refer to that manual.

Auditory Impaired

General Modifications

1. Each test item was carefully demonstrated.

2. Whenever possible, instructions to auditory impaired subjects were given by a person skilled in non-verbal communication.

3. When necessary, instructions were given in writing prior to testing.

4. All starting and stopping signals were given by hand signals and a foot stamp.

Specific Modification

Modified stork test. To begin this test, subjects were instructed to close their eyes and begin balancing when they were touched by the instructor.

Visually Impaired

General Modifications

Subjects were allowed a slow practice trial or walk through of each test item so that they had a clear idea of the activity.

Specific Modifications

Mat creep. The only adaptation necessary for the visually impaired involved providing the subject with a verbal cue when he/she was at a point just past the pylon and should turn around to creep back to the starting line. This cue was provided by saying "turn around" when the subject was just past the pylon.

Shuttle run. Blind subjects ran the course under either of two conditions: (a) with a sighted or functionally able partially sighted partner who was able to run faster than the blind partner or (b) alone but with the benefit of a guide rope or wire extended along the course. When blind subjects got to one end of the run, they bent down and touched the floor then returned to the starting line and repeated the process. Partially sighted subjects used brightly colored wooden blocks whenever possible.

Softball throw. A tactual aid may have been placed on the ground so that visually impaired subjects knew the location of the throwing line. The tactual aid may have been a narrow board, a different surface (dirt-grass), or some other aid.

Orthopedically Impaired

According to the Project UNIQUE Training Manual, modifications for the orthopedically impaired subjects were based on four classifications: orthopedic-amputee, orthopedic congenital anomaly, orthopedic cerebral palsy, and orthopedic spinal neuromuscular conditions. These categories are defined in the definitions section of Chapter I, under orthopedically impaired.

In the cases of orthopedic cerebral palsy and orthopedic spinal neuromuscular conditions a pre-established classification system was used to describe modifications. The National Association of Sports for Cerebral Palsy 1979 Classification Guide was used for the orthopedic cerebral

palsied subjects. For the orthopedic spinal neuromuscular conditions, the National Wheelchair Athletic Association classifications for individuals with spinal neuromuscular conditions was used. These classification systems are not listed or described here and the reader is refered to the Project UNIQUE Manual for this listing and/or description.

Orthopedically Impaired Amputees

Skinfold Measurements

One arm involvement. The triceps measure was taken on the nonimpaired arm.

Two arm involvement. These subjects were not tested on the triceps measure if involvement was above the elbow.

Two leg involvement. Some subjects needed to be supported for the subscapular and abdominal skinfolds as these measures were taken from a standing position.

Rise to Stand

Two leg involvement. Subjects did not participate in this activity.

Mat Creep

Two arm involvement. Subjects did not participate.

Two leg involvement. Subjects did not participate.

Shuttle Run

Two arm involvement. Subjects were instructed to bend down so that one knee touched the running surface. When the subject returned back to the starting line, he/she bent down and touched the surface again with a knee. This was repeated

until the shuttle run distance was completed.

One leg, two leg involvement or one arm-one leg involvement. This event was to be completed under one of the following conditions: move wheelchair forward with arms/feet move wheelchair backward with feet, cane, crutches, or no assistive device. If a wheelchair was utilized, the subject wheeled to the wooden blocks which were set up on a waste-basket tipped upside down (size ranging from 15" to 25" in height). The subject picked up one block and placed it on his/her lap and wheeled back to the starting line. Dropped the block down to the floor behind the starting line. Returned to the second block, picked it up off the basket, and placed it in the lap. Wheeled as quickly as possible past the starting/finish line to complete the shuttle run.

Modified Stork Test

One arm involvement. Subjects were required to cross one arm across their chest and cross any portion of the impaired arm.

Two arm involvement. Subjects were required to cross their chest with any portion of impaired limbs they possessed.

Two leg involvement. Subjects did not participate in this activity.

Sit and Reach Test

Two arm involvement. Subjects did not participate.

Two leg involvement. Subjects did not participate.

Grip Strength

One arm involvement. Subjects performed the test with

only the nonimpaired arm.

Two arm involvement. Subjects did not participate.

Flexed Arm Hang

One arm and two arm involvement. Subjects were not tested on this item.

Standing Broad Jump

One leg involvement. If the subject requested arm support during the jump, this assistance was provided.

Two leg involvement. Subjects did not participate.

Softball Throw

Two arm involvement. Subjects were not tested.

Two leg involvement. Subjects performed this activity from a seated position in a straight back or wheelchair. Subjects were given adequate warm up and two practice trials. Subjects must have brakes locked if using a wheelchair.

Orthopedically Impaired Congenital Anomalies

The same modifications described under orthopedically impaired amputees may be used for orthopedically impaired congenital anomalies. Some minor exceptions are listed below.

Rise to Stand

Two arm involvement. Subjects did not participate.

Shuttle Run

One leg involvement or one arm-one leg involvement. Some subjects did not participate in this activity. It was only administered to subjects who were capable of completing

the event in a reasonable period of time (30 seconds or less).

Two leg involvement. This item was administered only to subjects who used a wheelchair for daily activities. Then the same modifications described under orthopedically impaired amputees were used.

Orthopedically Impaired Cerebral Palsy

Rise to Stand

Classes I-IV. Subjects were not tested on this item.

Class VA. If necessary, a chair was used to aid the subject during stand.

Class VB. This item was not administered to Class VB subjects.

Mat Creep

Class I. Subjects did not participate in this event.

Shuttle Run

Class I. Subjects were not tested on this item.

Class II-IV. Subjects completed this task in a wheelchair. They propelled their wheelchair forward or backward using their feet or propelled their wheelchair forward with the arms. Subjects wheeled to the wooden blocks which are set up on a wastebasket tipped upside down (size ranging from 15" to 25" in height). The subject picked up one block and placed it on his/her lap and wheeled back to the starting line. Dropped the block to the floor behind the starting line. Returned to the second block, picked it up off the basket, and placed it in the lap. Wheeled as quickly as

possible past the start/finish line to complete the shuttle run.

Classes VA and VB. Instead of picking up blocks from the floor, subjects picked up blocks from a wastebasket tipped upside down (size ranging from 15" to 25" in height). The subject picked up one block and ran back to the starting line, dropped the block down to the floor behind the starting line, returned to the second block, picked it up off the basket, ran as quickly as possible past the starting/finish line to complete the shuttle run.

Modified Stork Test

Classes I-VB. Subjects were not tested on this event.

Sit and Reach

Classes I-III and Class VB. Subjects did not participate.

Grip Strength

Class I. Subjects were not tested.

Classes II-VII. Only limbs with functional strength were tested.

Flexed Arm Hang

Classes I-III. Subjects did not participate.

Classes IV-VII. Administered without modification to the fullest possible extent. Where arm involvement prohibited grasping, the bearing of weight, or reasonable execution, this item was omitted.

Standing Broad Jump

Classes I-IV and Class VB. Subjects were not tested.

Class VA. Administered without major modification. If

the subject requested arm support during the jump, this assistance was provided.

Softball Throw

Class I. Subjects were not tested.

Classes II-IV. Subjects performed the activity seated in a wheelchair.

Class VA-VB. Subjects were permitted to use a chair for stabilization during the softball throw.

Orthopedically Impaired Spinal Neuromuscular Conditions

Skinfold Measurements

Quadriplegic and paraplegic involvement. Some subjects needed to be supported while these measurements were taken.

Rise to Stand

Quadriplegic and paraplegic involvement. Subjects were not tested.

Mat Creep

Quadriplegic and paraplegic involvement. Subjects did not participate.

Shuttle Run

Quadriplegic and paraplegic involvement. Subjects using a wheelchair for daily activities participated in the shuttle run by utilizing a wheelchair. The subject wheeled to the two wooden blocks which are set up on a wastebasket tipped upside down (size ranging from 15" to 25" in height). The subject picked up one block and placed it on his/her lap and wheeled back to the starting line. Dropped the block down to the

floor behind the starting line. Returned to the second block, picked it up off the basket, and placed it in the lap. Wheeled as quickly as possible past the starting/ finish line to complete the shuttle run.

Modified Stork Test

Quadriplegic and paraplegic involvement. Subjects were not tested.

Sit and Reach

Quadriplegic and paraplegic involvement. Subjects did not participate.

Grip Strength

Quadriplegic involvement. Subjects were not tested.

Paraplegic involvement. Subjects were required to be seated for this task.

Flexed Arm Hang

Quadriplegic involvement. Subjects did not participate.

Paraplegic involvement. Subjects were helped to insure proper position. Careful spotting was employed.

Standing Broad Jump

Quadriplegic and paraplegic involvement. Subjects were not tested.

Softball Throw

Quadriplegic and paraplegic involvement. Subjects were tested on the softball throw from a seated position. Subjects were given adequate warm up and two practice trials. Subjects in wheelchairs had their brakes locked.