A RELIABILITY ANALYSIS OF THE AMERICAN
ASSOCIATION FOR HEALTH, PHYSICAL
EDUCATION AND RECREATION
YOUTH FITNESS TEST ITEMS

By

Arthur E. J. Field

B.P.E., The University of British Columbia, 1960

A Thesis Submitted in Partial Fulfilment of
The Requirements For The Degree of
Master of Physical Education
in the School
of
Physical Education
and
Recreation

We accept this thesis as conforming to the
required standard.

The University of British Columbia
August, 1964
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of PHYSICAL EDUCATION AND RECREATION

The University of British Columbia,
Vancouver 8, Canada

Date AUGUST 1964
ABSTRACT

A complete reliability analysis of the AAHPER Test has not been reported in the physical education literature. Previous reports have dealt only with the test-retest reliability coefficients of one or more items. The purpose of this study was to provide a comprehensive reliability analysis of the AAHPER Test items.

More specifically the problems of this study were (1) to determine the average and best test-retest reliability coefficients of the test items; (2) to determine the standard error of measurement (absolute accuracy) of the test items as computed by the standard correlation formula method and the analysis of variance technique; (3) to determine if the practice effect is significant for each test item; (4) to determine if the test items measure with an accuracy sufficient to distinguish between the subjects tested; and (5) to determine for each test item if subject differences (differences between subjects) are significantly larger than practice differences (differences between trials).

Fifty-seven untrained male students enrolled in the Required Physical Education Programme at the University of British Columbia were tested once a week for four consecutive weeks with the AAHPER Test. The items administered were pull-ups, sit-ups, standing broad jump, shuttle run, 50-yard dash, softball throw and 600-yard run-walk.

The data from each test item were analyzed in order to obtain (a) means and standard deviations for each of four trials, (b) between trials correlation coefficients and an average reliability coefficient, (c) standard errors of measurement computed by the standard correlation formula.
method and the analysis of variance technique, and (d) three F ratios (analysis of variance).

It was concluded on the basis of the reliability analysis of the data collected that (1) the average test-retest reliability coefficients of the test items were pull-ups .938, sit-ups .861, standing broad jump .899, shuttle run .776, 50-yard dash .792, softball throw .940 and 600-yard run-walk .759; (2) the standard errors of measurement computed by the standard correlation formula method and the analysis of variance technique were pull-ups (correlation formula method 0.794 and analysis of variance technique 0.834), sit-ups (6.250 and 6.934), standing broad jump (3.124 and 3.353 inches), shuttle run (0.227 and 0.239 seconds), 50-yard dash (0.194 and 0.190 seconds), softball throw (9.100 and 9.170 feet), and 600-yard run-walk (5.000 and 5.660 seconds); (3) analysis of variance results showed a significant practice effect over four trials for all items except the softball throw; (4) analysis of variance results showed that the AAHPER Test items measure with an accuracy sufficient to distinguish between the subjects tested; and (5) analysis of variance results showed that for each test item subject differences are not significantly larger than practice differences and since they usually are, it can be concluded that the practice effect must have been severe.

The findings of this study showed that the pull-ups and softball throw variables were highly reliable. Thus when using these items it seems reasonable to accept first trial scores as sufficiently accurate for both survey and experimental purposes. The standing broad jump, 50-yard dash and the 600-yard run-walk items had relatively high reliability; however, results showed that several preliminary practice trials are probably necessary before
scores become sufficiently reliable for research purposes. The sit-ups and shuttle run were the least reliable items of the AAHPER Test. These items seem to require at least four preliminary practice trials before a satisfactory level of reliability can be attained.
ACKNOWLEDGMENT

The writer wishes to thank Dr. S. R. Brown, of the University of British Columbia, School of Physical Education and Recreation for his stimulating advice and continued assistance in the writing of this study.

Sincere thanks are extended to the members of my thesis committee for their guidance and encouragement; to Mr. Harry Walters who aided in the administration of the tests; to Mr. R. W. B. Jackson, Statistician, University of Toronto and to Mr. H. Dempster and the staff of the University of British Columbia computer center.

Appreciation is extended to those fifty-seven students who served as subjects and produced the necessary data which made this study possible.

Finally, the writer wishes to thank his wife Joan and Mrs. E. Kane for their help in completing this study.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>STATEMENT OF THE PROBLEM</td>
</tr>
<tr>
<td></td>
<td>The Problem</td>
</tr>
<tr>
<td></td>
<td>Delimitations</td>
</tr>
<tr>
<td></td>
<td>Limitations</td>
</tr>
<tr>
<td>II</td>
<td>JUSTIFICATION OF THE PROBLEM</td>
</tr>
<tr>
<td></td>
<td>Importance of the Study</td>
</tr>
<tr>
<td></td>
<td>Definitions of Terms Used</td>
</tr>
<tr>
<td>III</td>
<td>REVIEW OF LITERATURE</td>
</tr>
<tr>
<td></td>
<td>History and Development of the AAHPER Test</td>
</tr>
<tr>
<td></td>
<td>Advantages of the AAHPER Test</td>
</tr>
<tr>
<td></td>
<td>Factors That Influence Test Reliability</td>
</tr>
<tr>
<td></td>
<td>Methods of Estimating Relative Precision (Reliability)</td>
</tr>
<tr>
<td></td>
<td>Methods of Estimating Absolute Precision (Errors of Measurement)</td>
</tr>
<tr>
<td></td>
<td>Related Reliability Studies</td>
</tr>
<tr>
<td>IV</td>
<td>METHODS AND PROCEDURE</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
</tr>
<tr>
<td></td>
<td>Test Administration</td>
</tr>
<tr>
<td></td>
<td>Testing Team</td>
</tr>
<tr>
<td></td>
<td>Equipment and Facilities</td>
</tr>
<tr>
<td></td>
<td>Statistical Treatment of Data</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>PAGE</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>V RESULTS AND DISCUSSION</td>
<td>36</td>
</tr>
<tr>
<td>Reliability</td>
<td>36</td>
</tr>
<tr>
<td>Standard Errors of Measurement</td>
<td>41</td>
</tr>
<tr>
<td>Analysis of Variance</td>
<td>42</td>
</tr>
<tr>
<td>VI SUMMARY AND CONCLUSIONS</td>
<td>51</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>54</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>59</td>
</tr>
<tr>
<td>A. AAHPER RESEARCH PROGRAMME</td>
<td>60</td>
</tr>
<tr>
<td>B. HISTORY AND USE OF THE AAHPER TEST</td>
<td>63</td>
</tr>
<tr>
<td>C. AAHPER TEST ADMINISTRATION INSTRUCTIONS</td>
<td>65</td>
</tr>
<tr>
<td>D. AAHPER TEST SCORE CARD</td>
<td>69</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
</tr>
<tr>
<td>I AAHPER Youth Fitness Test Data From Four Trials</td>
<td>37</td>
</tr>
<tr>
<td>II AAHPER Youth Fitness Test Standard Errors of Measurement</td>
<td>41</td>
</tr>
<tr>
<td>III AAHPER Youth Fitness Test Analysis of Variance Data</td>
<td>43</td>
</tr>
</tbody>
</table>
CHAPTER I

STATEMENT OF THE PROBLEM

If we are to attach any significance to the results of the American Association for Health, Physical Education and Recreation (AAHPER) Youth Fitness Test, we must first be assured that the test items are accurate and reliable measuring instruments. We can not justify the use of the AAHPER Test as a measuring instrument of youth fitness if the test items are inaccurate and unreliable.

The AAHPER Youth Fitness Test was developed by the American Association for Health, Physical Education and Recreation to test the physical fitness of the American youth. It is now being widely used in schools and universities throughout the world. It came into existence to supply the need for a suitable test to be used in conjunction with the "Fitness Movement" in the United States. This movement was initiated by President Eisenhower, and continued by the late President Kennedy's Council on Youth Fitness.

The AAHPER National Youth Fitness Test provides national norms of physical performance for boys and girls, ages 10-17 and college age youth and young adults, ages 18-30. The test is especially useful for determining individual weaknesses, program quality and changes in physical performance.

The items of this test are pull-ups on the horizontal bar for measuring arm strength, sit-ups for measuring abdominal endurance, shuttle run for measuring speed and agility, standing broad jump for measuring muscular power of the legs, 50-yard dash for determining speed, softball
throw for measuring arm power, and the 600-yard run-walk for determining endurance during sustained activity. The test may be administered in a gymnasium or outside without special equipment.

I. THE PROBLEM

The object of this study is to determine by a reliability analysis if the AAHPER Youth Fitness Test items are accurate and reliable measuring instruments. More specifically the problems of this study were:

1. to determine the average and best test-retest reliability coefficients of the test items;
2. to determine the standard errors of measurement (absolute accuracy) of the test items as computed by the standard correlation formula method and the analysis of variance technique;
3. to determine if practice effect is significant for each test item;
4. to determine if the test items measure with an accuracy sufficient to distinguish between the subjects tested; and
5. to determine for each test item if subject differences (differences between subjects) are significantly larger than practice differences (differences between trials).

II. DELIMITATIONS

1. This study deals with a group of fifty-seven untrained male first and second year University of British Columbia students.
III. LIMITATIONS

1. A larger sample would be more desirable in this type of investigation.

2. The weather conditions during the second outdoor session were not ideal.

3. Some subjects did not participate to the best of their ability.

4. No attempt was made to obtain a random sample. All subjects volunteered for the testing program.
CHAPTER II

JUSTIFICATION OF THE PROBLEM

The function of measurement in physical education is to assess status or capacity in a given quality or skill at a given time.\(^1\) In order to attain the goals of physical education it is necessary to use measuring devices such as the AAHPER Test. The AAHPER Youth Fitness Test is one of the latest, most valid physical fitness tests to be constructed. The supposition is that this test is an accurate and reliable measuring instrument of the most important aspects of physical fitness.

Reliability coefficients of the individual items included in the American Association for Health, Physical Education and Recreation Youth Fitness Test have been reported from time to time by various investigators. No previous report appears to have been made of the reliability of the AAHPER Test items when done in successive order by the same subjects under conditions which have been officially recommended for the administration of the whole test. The test-retest reliability coefficients reported by various other investigators describe only one aspect of reliability. Other aspects of reliability are equally or more important and have been generally ignored by test designers and research workers in physical education. Moreover, most reports of reliability using correlation coefficients have been based on only two trials. The possibility of further improvement in reliability by repeated testing has thus been virtually ignored. The object of this study was to determine

by a comprehensive reliability analysis if the American Association for Health, Physical Education and Recreation Youth Fitness Test items are accurate and reliable measuring instruments.

I. IMPORTANCE OF THE STUDY

Complete and detailed information concerning the reliability of the AAHPER Test as a measuring device of youth fitness is of considerable importance in the field of physical education. This information will provide valuable assistance for the physical educator in determining the test's usefulness when dealing with a specific problem and in interpreting test results.

In discussing the importance of reliability in educational measurement, Thorndike stresses the following important concepts which are applicable to measurement and evaluation in physical education.

In physical education we obtain a score for an individual on a test in order to arrive at some conclusion concerning him or to map a program of action with regard to him. When selecting a physical performance test for a specific testing project and when interpreting the test results, the physical educator is primarily concerned with the reliability and accuracy

---


of the measuring instrument. Assuming all other considerations to be equal (particularly validity) the physical educator will choose that test which will give the most accurate and reliable estimate of the characteristic being studied.

An individual score of questionable reliability resulting from the application of a physical performance test is distressing to the research worker in physical education. It would be unwise to consider this score as an indication of true individual performance. Further action or judgement which may be based on an unreliable score can only be tentative.

II. DEFINITIONS OF TERMS USED

Reliability

It is apparent that there is some disagreement among the authorities as to the definition of reliability. This disagreement stems in part from an early connection between reliability and the correlation coefficient which has tended to confuse rather than clarify the issue. This confusion can best be explained by examining some of the definitions of reliability which are in common use.

Most writers in physical education define reliability as consistency. Clarke defines reliability as "... the degree of consistency with which a

4 Jackson and Ferguson, Reliability of Tests, p. 23.
measuring device may be applied."\(^5\) Weiss and Scott give a similar
definition of reliability. They define reliability as the "...consistency with which a test can be administered by the same tester."\(^6\)Meyers and Blesh in agreement with Clarke and Weiss and Scott summarize by defining reliability as "... the degree of consistency of results obtained by a measuring instrument upon repeated application..."\(^7\)

Authorities in educational research give additional definitions of reliability. Thorndike and Hagen state that "Reliability has to do with accuracy and precision of a measurement procedure."\(^8\) Kientzle defines reliability of a test as "... the extent to which it agrees with itself."\(^9\) Garrett states "A test score is called reliable when we have reasons for


\(^7\) Meyers and Blesh, Measurement in Physical Education, p. 62.


believing the score to be stable and trustworthy." Guilford reports that "Under the concept of reliability we are concerned about the accuracy with which a score represents the status of an individual in whatever aspect the test measures him."\(^{11}\)

Guilford also points out that reliability has several meanings. "Common synonyms for reliability include: dependability, consistency, and stability. Each means something somewhat different as applied to measurements. Even the same term has slightly different meanings as applied to different measurement operations."\(^{12}\)

Jackson and Ferguson\(^{13}\) state that much of the disagreement among the various writers as to the definition of reliability stems from the fact that some authorities place emphasis on accuracy or errors of measurement while other authorities place emphasis on consistency or relative stability (reliability). They suggest that the issue would be


clarified if the two terms absolute accuracy and relative accuracy were used in the place of the blanket term reliability.

Guilliksen\textsuperscript{14} reports that it has become customary in the last forty years to assess tests in terms of the reliability coefficient rather than the error of measurement. He suggests in agreement with Jackson and Ferguson that since the reliability coefficient and the standard error of measurement have advantages and disadvantages, both should always be given when making a complete assessment of a test. Thorndike\textsuperscript{15} in collaboration with Cronbach, Cureton, Kelly, Kurtz, Richardson, and Thurstone defines reliability as the tendency towards consistency. He states in agreement with Jackson and Ferguson and Guilliksen that the consistency of a set of measures may be determined by two different methods. In the first, one is concerned with the amount of variation we would find in a set of repeated measurements of the same specimen. This is called the standard error of measurement. The second method of determining consistency in measurement may be made in terms of the consistency with which an individual maintains his ranking in a group when a set of measures are reproduced a second time. An index of this consistency of measurements is the correlation between two sets of scores which may be called the reliability coefficient.


Lyman\textsuperscript{16} in agreement with Jackson and Ferguson, Guilliksen, and Thorndike reports that reliability refers to consistency of measurement and may be assessed in either an absolute or relative sense. He states absolute consistency refers to the variability we could expect in a person's score, if examined repeatedly with the same test while relative consistency refers to the ability of the test to reproduce scores that place examinees in the same position relative to each other.

It appears that reliability is a somewhat difficult term to define. This difficulty stems in part from the fact that reliability refers to a series of concepts that are often confused with one another. Authorities, however, seem to agree that any manual that accompanies a test should provide as statistical indices of precision of measurement, measures expressed both in relative (correlation coefficient) and absolute (standard error of measurement) terms.

**Standard Error of Measurement**

The standard error of measurement indicates how much we would expect a person's score to vary if he were examined repeatedly with the same test. It expresses the reliability of a test in an absolute sense. This statistic estimates how accurate our obtained scores are and how much confidence we can place on getting a reasonably accurate measurement with

the variable concerned. Thus one can have 68.26 percent confidence the obtained score lies within \( \pm 1 \) standard error of measurement of the true score or 99 percent confidence that the obtained score lies within \( \pm 2.58 \) standard errors of measurement of the true score.

**Practice Effect**

The practice effect refers to the improvement in performance from test to retest. The practice effect is also called the initial educational adjustment and may be attributed to learning rather than to physiological changes. Factors which contribute to the practice effect are increased motivation, competition with other subjects, improvement in form, preliminary practice, coaching, and changes in interest, desire and attitude.
CHAPTER III

REVIEW OF LITERATURE

Much has been written on the American Association for Health, Physical Education and Recreation Youth Fitness Test and the concept of reliability, however, this review will present only a brief summary of the AAHPER Test and a survey of the literature closely related to the problem at hand.

I. HISTORY AND DEVELOPMENT

OF THE AAHPER TEST

During the war years there was a great interest in fitness but during times of peace this interest fell off. Much of the interest during the last decade can no doubt be traced to the large draftee rejection rate of the Korean War.

The results of the Kraus-Weber Test did much to revitalize interest in fitness. In 1946 Dr. Hans Kraus, Associate Professor at the Institute of Rehabilitation, New York University, developed the Kraus-Weber Test for Minimum Muscular Fitness. The Kraus-Weber Test is composed of five strength items and one flexibility item. Each movement is performed once on a pass or fail basis.

The Kraus-Weber Test was originally designed to determine if polio patients were sufficiently rehabilitated to leave the hospital. The test was later applied to posture cases and as a measure of progress for those afflicted with disorders of the spine.
The tests were later administered to 4,264 American and 2,870 European children between the ages of 6 and 16 by Kraus and Hirschland. Results showed that 57.9 percent of the American children and only 8.7 percent of the European children failed.

Immediately following the release of the Kraus-Weber Test results, leading newspapers and magazines carried articles pertaining to the fitness of America's youth. These articles aroused the Nation's interest in youth and adult fitness. Consequently, individuals began to ask: what is wrong with our fitness; what can be done and what must be done?

This great interest in fitness led former President Eisenhower to call a conference on the fitness of American youth. This conference, the first fitness conference held under the auspices of the Federal Government, was held at the United States Naval Academy, Annapolis, June 18-19, 1956.

The Conference recommended that:

Official recognition be taken of the fact that our adult citizens and our youth have little appreciation of the existence of a problem pertaining to the fitness of American youth.

The public generally, and parents, church leaders, and educators in particular, be alerted to the facts that (a) in this age of automation, the fitness of our youth cannot be taken for granted, (b) indifference to the softness which comes from lack of participation in

---

health-giving activities will bring erosion of our strength, and (c) physical fitness goes hand-in-hand with moral, mental, and emotional fitness.

Intensive, continual, and co-operative research be conducted to supply the factual base for formulating fitness policies, plans, and programs.2

As a result of this conference President Eisenhower created the President's Council of Youth Fitness on September 6th, 1956.

The results of the Kraus-Weber Tests and the findings of the National Conference on Youth Fitness prompted the American Association for Health, Physical Education and Recreation to hold a conference on September 12-13. It was decided at this conference that a meeting of selected members of the Research Council would be held in February 1957 in order to formulate a National Fitness Project.

The initial planning of the AAHPER Test took place at this follow-up meeting in Chicago. At the end of two day's discussion the test battery was decided upon. The test included items which attempted to judge the individual's proficiency in running, throwing, endurance, strength, agility and swimming. Specifically, the items were pull-ups for measuring arm-shoulder girdle strength; sit-ups for measuring strength and endurance of abdominal and hip flexor muscles; shuttle run for measuring speed, co-ordination and agility, standing broad jump for measuring explosive power

of the leg extensors; 50-yard dash for measuring speed, softball throw for measuring skill, strength and co-ordination; and the 600-yard run-walk for measuring endurance.

Shortly after the Chicago meeting a Physical Fitness Research Committee was formed with Paul Hunsicker as director. Under the direction of Dr. Hunsicker the AAHPER Test was administered to a total of 8,500 school children in grades 5 to 12. The results were analyzed and national norms were established. In addition, a test manual was prepared by the American Association for Health, Physical Education and Recreation describing the test battery and its administration.

II. ADVANTAGES OF THE AAHPER TEST

Hunsicker\(^3\) reports five reasons why items of the AAHPER Test were chosen by the Advisory Committee of the AAHPER Youth Fitness Testing Project.

1. The Tests are Reasonably Familiar

The AAHPER Youth Fitness Test items have been used for some time and are familiar to, or readily understood by all physical education personnel. Even young boys and girls and those individuals without

specialized training in physical education should have little or no difficulty understanding the test items.

2. The Tests Require Little or No Equipment

The equipment required to administer the AAHPER Test is of a low cost and is usually found in most schools and recreation areas.

3. The Tests Can Be Administered to Both Sexes

The test items can be administered to both boys and girls; however, the pull-up item is modified for girls.

4. The Tests Can Be Administered to Grades 5 to 12

The tests can easily be given to children in the age range of grades 5 to 12.

5. The Tests Measure Different Aspects of Fitness

The test items measure the different components of fitness and many muscle groups of the body.

III. FACTORS THAT INFLUENCE TEST RELIABILITY

The reliability of a test is influenced by a wide variety of causes. The more common factors that affect test reliability are listed below:
1. **Increasing Test Length**

Increasing the length of a test will increase reliability providing of course the group is the same and the new items are as good as those on the shorter test.

2. **Repeating a Test**

Averaging the scores from several applications of a test or from parallel forms will increase reliability.

3. **Variability of the Group**

The size of variance is important in determining reliability. A wide distribution of group scores is more likely to yield a higher reliability coefficient than a narrow restricted distribution.

4. **Time**

In a test-retest situation the reliability is higher when the time interval between the two testing periods is short.

5. **Item Difficulty**

Reliability is higher when test items are about fifty percent difficult. Reliability tends to decrease when the difficulty of the items departs from the fifty percent level.

6. **Practice**

Any factor such as practice which influences the difficulty
value of the test items will also influence reliability.

7. **Irregularities**

Irregular testing conditions which affect some scores more than others tend to cause lower reliability coefficients.

8. **Item Intercorrelations**

Reliability is highest where the items of the test all intercorrelate highly.

9. **Range of Difficulty**

The more nearly equal are the difficulties of the test items, the higher the reliability.

**IV. METHODS OF ESTIMATING RELATIVE PRECISION (RELIABILITY)**

There are five procedures in use for computing the reliability of a test. These methods are as follows:

1. **Alternate or Parallel Forms Method (Equivalent Forms)**

When duplicate forms of a test are administered to the same group of subjects, the correlation between Form A and Form B may be used as the reliability of the test.

When developing two parallel forms of a test care must be taken to ensure similar test content and administration instructions. The parallel
forms method is the most commonly used estimate of reliability for written standardized tests.4

2. Test-Retest Method (Stability)

When a test is administered to the same group of subjects on two occasions, the correlation between the scores earned on the two administrations may be used as the reliability of the test. The test-retest method is most often used when determining the reliability of skill and physical performance tests.

Thorndike states that "... for those types of tests in which sampling of items and memory of previous responses are not an issue and for which comparability of motivation seems likely, a second application of the same test at a later date, and correlation of the two sets of scores provides an adequate set of operations for reliability estimation."5

Garrett points out that "The test-retest method will estimate less accurately the reliability of a test which contains novel features and is highly susceptible to practice than it will estimate the reliability of test scores which involve familiar and well-learned operations little affected by practice. Owing to difficulties in controlling conditions which


influence scores on retest, the test-retest method is generally less useful than other methods.\(^6\)

3. **Split-Half Method (Internal Consistency)**

In this method of determining reliability the test is divided into two equivalent halves. Scores from the two halves are then correlated. The correlation obtained, however, represents the reliability coefficient of only one half the test, so to obtain the reliability of the whole test the Spearman-Brown Prophecy Formula must be used. This formula is:

\[
\rho_{XX} = \frac{2\rho_{hh}}{1 + \rho_{hh}}
\]

where \(\rho_{hh}\) is the reliability of a half test.

The question of how a test may be divided into two halves is an important one. The procedures include:

(a) Putting alternate items in each half test (odd-even split).
(b) Using the first half of the test as one-half test and using the second half as the other (first versus second halves).
(c) Selecting items for the two half tests which are equivalent in content and difficulty.
(d) Putting alternate groups of test items in each half-test.

The split-half method of determining reliability is used when it is not possible to repeat the test or to construct parallel forms of the test.

---

One of the advantages of the split-half method is that all the data required for computing reliability can be obtained at one sitting.

4. Method of Rational Equivalence

The formulas used in this method of calculating reliability are somewhat different from those used in the calculation of reliability discussed so far. The method of rational equivalence does not require the calculation of a correlation coefficient. This method determines the internal consistency of a test through an analysis of the individual test items. The formulas used to calculate this method of reliability are known as the Kuder-Richardson formulas. Garrett gives a simple approximation of one of the Kuder-Richardson formulas which is useful in determining reliability quickly. It reads

$$\gamma = \frac{N \sigma_t^2 - M(N - M)}{\sigma_t^2(N - 1)}$$

Where:

- $\gamma$ = reliability of the whole test
- $N$ = number of items on the test
- $\sigma_t$ = standard deviation of the test scores
- $M$ = the mean of the test scores

The Kuder-Richardson rational equivalence formulas give a lower estimate of test reliability than would be obtained by the other methods.

---

described. Thus, using this method of determining reliability eliminates the danger of making an overestimation.

5. **Analysis of Variance Methods**

Hoyt\(^8\) has developed a formula for estimating test reliability by analysis of variance techniques. He showed that the reliability of a test could be estimated from the formula

\[
\gamma = \frac{S^2}{(K-1)} - \frac{S^0}{(K-1)(N-1)}
\]

Where:

\[
\frac{S^2}{K-1} = \text{among individuals mean square (variance)}
\]

\[
\frac{S^0}{(K-1)(N-1)} = \text{error mean square (variance)}
\]

Jackson and Ferguson\(^9\) applied analysis of variance techniques and the methods of testing statistical hypotheses to estimate the reliability of two forms of an intelligence test. To determine the relative accuracy they introduced a new statistic called the sensitivity coefficient. It is defined as the ratio of the standard deviation of the true scores to the

---


standard deviation of the distribution of errors of measurement. In other words the sensitivity coefficient expresses the differences between the individuals tested in terms of the errors of measurement of the test. If the sensitivity coefficient is small, then the errors of the measurement will be large in comparison with the differences in individual ability tested and hence the scores obtained by an individual will have a large error component.

To find a unique estimate of the sensitivity (relative accuracy) from an analysis of variance proceed as follows:

(a) Subtract the error mean square from the between subjects mean square.
(b) Divide the resulting difference by twice the error mean square.
(c) Find the square root of the quotient as an estimate of the sensitivity.

Alexander\(^{10}\) using the methods of estimation provided by analysis of variance, provides estimates of reliability for a block of data consisting of several trials by the same individuals. He also discussed the conditions under which each estimation of reliability is considered valid.

V. METHODS OF ESTIMATING ABSOLUTE PRECISION
(ERRORS OF MEASUREMENT)

Reliability can be expressed in either absolute or relative terms. This became evident when reviewing the various definitions of reliability. Absolute precision takes the form of the standard error of measurement. The standard error of measurement is the standard deviation of a distribution of scores all of which are estimates of the same true score. The standard error of measurement may be estimated by the standard correlation formula method or the analysis of variance technique.

1. **Standard Formula Method**

The reliability coefficient by itself does not give an estimate of absolute accuracy of our measurements. However, we can calculate the standard error of measurement using the reliability coefficient in the formula

\[ SE_M = SD \sqrt{1 - r} \]

Where:  
- \( SE_M \) = the standard error of measurement  
- \( SD \) = the standard deviation of the test scores (test I)  
- \( r \) = the reliability coefficient of test I

2. **Analysis of Variance Technique**

The standard error of measurement may also be calculated by analysis of variance by taking the square root of the error mean square. This gives a direct estimate of the absolute accuracy of our measurements.
VI. RELATED RELIABILITY STUDIES

A review of the literature was made to discover the reported reliability of the AAHPER Test items.

Barrow\textsuperscript{11} in an attempt to develop an easily administered test of general motor ability for college men calculated the test-retest reliability of twenty-nine test items. He found the following test-retest reliabilities for items similar to those used on the AAHPER Test; standing broad jump .895, softball throw .928, and 60-yard dash .828.

Kane and Meredith\textsuperscript{12} in determining ability in the standing broad jump of elementary school children 7, 9, and 11 years of age, calculated reliability coefficients for boys and girls in each age group. Pearson product moment correlation coefficients were computed for the best record with the second best record of twelve trials. For all three ages and both sexes, the reliability coefficients approximated .980.

In 1952, McCraw and Tolbert\textsuperscript{13} conducted a study comparing different


\textsuperscript{12} Robert J. Kane and Howard V. Meredith, "Ability in the Standing Broad Jump of Elementary School Children 7, 9, and 11 years of Age," \textit{AAHPER Research Quarterly}, Vol. 23 (May 1952), pp. 198-208.

methods of scoring, such as, one trial, average of three trials, median of three trials, best one of three trials, average of two trials and best one of two trials. Six tests of physical ability were administered to 128 junior high school boys. The tests of physical ability used were 50-yard dash, standing broad jump, softball throw, jump and reach, a wall volley and speed shooting. Subjects were allowed three trials for each test on each of two administrations. The following coefficients of correlation were reported.

<table>
<thead>
<tr>
<th>Trials</th>
<th>50-Yard Dash</th>
<th>Standing Broad Jump</th>
<th>Softball Throw</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Trial - 2nd Trial, First Adm.</td>
<td>.886</td>
<td>.818</td>
<td>.933</td>
</tr>
<tr>
<td>1st Trial - 3rd Trial, First Adm.</td>
<td>.845</td>
<td>.831</td>
<td>.914</td>
</tr>
<tr>
<td>2nd Trial - 3rd Trial, First Adm.</td>
<td>.905</td>
<td>.858</td>
<td>.908</td>
</tr>
<tr>
<td>1st Trial - 2nd Trial, Second Adm.</td>
<td>.855</td>
<td>.902</td>
<td>.930</td>
</tr>
<tr>
<td>1st Trial - 3rd Trial, Second Adm.</td>
<td>.882</td>
<td>.877</td>
<td>.920</td>
</tr>
<tr>
<td>2nd Trial - 3rd Trial, Second Adm.</td>
<td>.893</td>
<td>.913</td>
<td>.944</td>
</tr>
<tr>
<td>First Trial - First Trial</td>
<td>.832</td>
<td>.797</td>
<td>.892</td>
</tr>
<tr>
<td>Average of Three-Average of Three</td>
<td>.931</td>
<td>.918</td>
<td>.940</td>
</tr>
<tr>
<td>Median of Three - Median of Three</td>
<td>.861</td>
<td>.899</td>
<td>.905</td>
</tr>
<tr>
<td>Best of Three - Best of Three</td>
<td>.876</td>
<td>.916</td>
<td>.927</td>
</tr>
<tr>
<td>Average of Two - Average of Two</td>
<td>.886</td>
<td>.882</td>
<td>.905</td>
</tr>
<tr>
<td>Best of Two - Best of Two</td>
<td>.864</td>
<td>.845</td>
<td>.915</td>
</tr>
</tbody>
</table>

Brown revealed the reported range of reliability of twenty-eight

14

selected test items including six which are similar to the items on the AAHPER Test. He also determined the reliability of the test items himself.

<table>
<thead>
<tr>
<th>Test Items</th>
<th>Range of Previously Reported Reliabilities</th>
<th>Reliability Found By Brown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pull-Ups or Chins</td>
<td>.86-.99</td>
<td>.915</td>
</tr>
<tr>
<td>60-Yard Dash</td>
<td>.84-.97</td>
<td>.823</td>
</tr>
<tr>
<td>Standing Broad Jump</td>
<td>.66-.98</td>
<td>.905</td>
</tr>
<tr>
<td>Softball Throw</td>
<td>.89-.97</td>
<td>.916</td>
</tr>
<tr>
<td>Sit-Ups (2 minutes)</td>
<td>.71-.99</td>
<td>.823</td>
</tr>
<tr>
<td>60-Yard Shuttle Run</td>
<td>no report</td>
<td>.819</td>
</tr>
<tr>
<td>Dodge Run</td>
<td>.68-.97</td>
<td>.875</td>
</tr>
</tbody>
</table>

Willgoose, Askew, and Askew estimated the reliability of the 600-yard run-walk at the grade eight level. They tested seventy grade eight boys at an interval of one week under identical testing conditions. A rank order correlation coefficient of .92 was computed.

CHAPTER IV

METHODS AND PROCEDURE

This chapter presents a detailed description of the methods and procedures used in administering and analyzing the seven items of the AAHPER Test.

I. SUBJECTS

One hundred and twenty University of British Columbia first and second year male students registered in the Physical Education Research Programme in the Fall of 1961. These students agreed to become subjects for a number of physical fitness tests to be administered by fourth year students, and graduate students in physical education. Registration in the Physical Education Research Programme enabled students to satisfy one-half of the Physical Education Service Programme requirement.

During the month of March, 1962, students who had enrolled in the Research Programme received a letter from Dr. Stanley R. Brown, Associate Professor and Director of Research, informing them of the testing procedure (see Appendix A).

Each subject was required to attend two testing sessions a week for four weeks. Testing began on Monday, March 12th at the University of British Columbia. Indoor sessions were held on Mondays and Wednesdays in the University of British Columbia War Memorial Gymnasium. Outdoor sessions were held Wednesdays and Fridays on the Varsity Stadium playing field.

One hundred and twenty letters were sent to students originally
enrolled in the Research Programme; however, only seventy-nine of these one hundred and twenty students began the testing programme. At the end of the fourth week fifty-seven subjects had completed the testing programme.

II. TEST ADMINISTRATION

Before the first testing session each subject was given a summary of the history and use of the AAHPER Test (see Appendix B) and a summary of test administration instructions (see Appendix C).

All subjects made four complete runs through the seven item American Association for Health, Physical Education and Recreation Youth Fitness Test. The seven items administered were pull-ups, sit-ups, standing broad jump, shuttle run, 50-yard dash, softball throw and 600-yard run-walk. Pull-ups, sit-ups, standing broad jump and shuttle run items were done indoors and 50-yard dash, softball throw, and 600-yard run-walk items were done out-of-doors. Sit-ups were done on a gymnasium wood floor, the 50-yard dash and softball throw were done on a grass football field and the 600-yard run-walk was conducted on a 440 yard cinder track.

Subjects were given a demonstration by their instructors of the proper method of performing the test items. They were given similar motivation throughout the experiment and were tested by the same group of testers each week. The testing instructions specified were carefully followed in all cases. For the purpose of this study it was necessary to emphasize before testing began that no warm-up or practice was allowed.

Each participant was given an AAHPER Youth Fitness Test score card
for recording results (see Appendix D). These cards were to be filled out after the completion of each test item and carried to the next testing station. Subjects were instructed to follow the pre-established order of performance until they completed the test items. Students worked in pairs to observe and record each other's performance under supervision.

Weather conditions were uniformly good during the first, third and fourth trials but during the second outdoor testing period there was a continuous light drizzle with no wind. As it was not possible to postpone the outdoor portion of the second testing period, it was considered worthwhile recording results to see what happens to performance and reliability measurements under conditions which are common to the west coast of British Columbia.

III. TESTING TEAM

The testing team was made up of graduate students and qualified physical education instructors. All testers took part in a training session before the first trial was administered. At this time, specific instructions were given, testing procedures were demonstrated and a practice session was conducted. All members of the testing team were familiar with the testing routine before the testing programme began.

IV. EQUIPMENT AND FACILITIES

The following equipment and facilities were used for the administration of the test.

1. Gymnasium area
2. Outdoor playing field
3. High bar
4. Two gymnasium mats
5. Four small blocks 2" x 2" x 4"
6. Two tape measures
7. Two stop watches
8. Six Voit rubber softballs
9. Football yard markers, one to three hundred yards.

V. STATISTICAL TREATMENT OF DATA

Four sets of data for each of the AAHPER Test items were gathered by the investigator. These data were punched on I.B.M. cards and were analyzed to obtain (1) test item means and standard deviations on each of four trials, (2) an intercorrelation matrix and (3) three F ratios (analysis of variance) for each test item.

Average Reliability Coefficients

In order to calculate an average reliability coefficient for each variable, the between trials reliability coefficients obtained from the intercorrelation matrix were transformed into Fisher's Z function, summed and divided by their number to obtain the arithmetic mean of the Z's. The mean Z was then converted into an equivalent reliability coefficient.

Standard Errors of Measurement

Standard errors of measurement (absolute accuracy) for each variable were calculated by the analysis of variance technique and the standard
correlation formula method.

The standard errors of measurement using the analysis of variance technique were calculated for each variable by taking the square root of the error mean square. The standard errors of measurement for each variable using the standard correlation method were calculated by incorporating the average standard deviation and the average reliability coefficient into the standard error of measurement formula

\[ SE_M = SD_A \sqrt{1 - \gamma_A} \] (where \( A \) = average). For the three variables (50-yard dash, softball throw and 600-yard run-walk) measured out-of-doors under wet conditions during the second testing period, standard errors of measurement were calculated with trial two data eliminated.

Analysis of Variance

To determine the usefulness of the AAHPER Test items as measuring instruments, the testing results for each test item were analyzed by analysis of variance techniques advocated by Jackson and Ferguson\(^1\) and Garrett.\(^2\)

The data from each of the outdoor test items (50-yard dash, softball throw and 600-yard run-walk) were analyzed by analysis of variance with trial

---

\(^1\) Robert W.B. Jackson and George A. Ferguson, Studies on the Reliability of Tests, University of Toronto, Toronto, Department of Educational Research, Bulletin No. 12, 1941, pp. 31-38.

two data eliminated. This procedure was followed because the means of the outdoor items were adversely affected by wet weather conditions which prevailed during the second outdoor testing session.

The analysis of variance methods used in this study presuppose that the subjects tested are not all of the same ability and that subject scores, on the average, progressively increase from trial to trial. From the data collected for each test item it was quite obvious that the subjects varied considerably in ability. In most cases, it also appeared that, on the average, subject scores improved from trial to trial. Where this apparent improvement was tested statistically and found to be significant it could properly be described as practice effect.

Analysis of variance techniques were used in this study for treating three different problems: (1) The determination of a significant practice effect, (2) the determination of whether or not the test items measure differences between subjects tested, and (3) the determination of whether or not subject differences are significantly greater than practice differences.

In order to investigate these problems it was necessary to obtain a measure of the amount the practice effect, the differences between subjects, and error contributed to the total variance. The total variance was divided into components which were assigned to practice effect (trials), differences between subjects (individuals), and error (residual). This provided a means by which to test the significance of the practice effect and the differences between subjects and to determine whether or not
subject differences are significantly greater than practice differences.

**Practice Effect (Differences Between Trials)**

In order to determine whether or not there was a significant practice effect for each test item, it was necessary to calculate an F ratio for trials. The F ratio for trials was calculated by dividing the mean square (variance) trials by the error mean square (variance).

If the practice effect is not significant (no significant differences between the means), the between trials mean square will be of the same order as the error mean square. However, if the between trials mean square is significantly greater than the error mean square, then it may be concluded that the practice effect is significant.

**Differences Between Subjects**

In order to determine if a test item measure with an accuracy sufficient to distinguish between subjects tested it was necessary to calculate an F ratio for subjects. The F ratio for subjects was calculated by dividing the mean square (variance) subjects by the error mean square (variance).

In most groups individuals will not be of the same ability in any test of motor performance. If the accuracy with which any test measures performance is not sufficient to distinguish between the abilities of the subjects tested, then the differences between the scores made by the subjects will be small and due entirely to errors of measurement. This means that the
between subjects mean square will not be significantly larger than the error mean square. However, if between subjects mean square is significantly larger than the error mean square, it may be assumed that the differences between subject scores are too large to be caused solely by errors of measurement. Thus, it may be concluded that the test measures performance with an accuracy sufficient to distinguish between the abilities of the subjects tested. In other words the test can be relied upon to produce consistent results.

Subject Differences Versus Practice Differences

Through a further analysis of variance it is possible to determine whether or not subject differences (differences between subjects) are significantly greater than practice differences (differences between trials). This method of analysis was used to calculate a third $F$ ratio for each variable. This ratio was obtained by dividing the mean square (variance) subjects by the mean square (variance) trials.

If the mean square subjects is not significantly larger than the mean square trials, it may be concluded that subject differences are no larger than practice differences. Subject differences are, however, usually much larger than practice differences, but when they are not, it may be assumed that there has been a practice effect.
CHAPTER V

RESULTS AND DISCUSSION

The results given in this study refer to the scores made by fifty-seven subjects on four repetitions of the seven item AAHPER Youth Fitness Test. Results are presented and discussed in terms of reliability, errors of measurement and analysis of variance (F ratios).

I. RELIABILITY

In experimental studies employing motor fitness (physical performance) items, it is desirable to repeat pre-training trials until satisfactory reliability standards are met. This seems necessary whether control groups are used or are not used since failure to show differences between experimental and control groups may be a result of data being unreliable. Also there is the practical question of how much improvement might have resulted solely from a practice of the test item. If practice alone can produce changes comparable with training, what real knowledge has been gained about the value of the training method in improving fitness?

Strictly reliable motor fitness data show (a) relatively high correlation coefficients, (b) means of similar size, (c) standard deviations of comparable size, (d) differences between paired scores which have the appearance of being random errors and (e) small standard errors of measurement.

Test item performance means and standard deviations for the four trials are presented in Table I together with between trials reliability coefficients and an average reliability coefficient.
### TABLE I

**AAHPER YOUTH FITNESS TEST**

**DATA FROM FOUR TRIALS**

*(N=57)*

<table>
<thead>
<tr>
<th>Variable and Trials</th>
<th>Mean Scores</th>
<th>Standard Deviations</th>
<th>Between Trials</th>
<th>Test-Retest Reliability Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Pull-Ups (No.)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6.245</td>
<td>3.465</td>
<td>1 v 2</td>
<td>.917</td>
</tr>
<tr>
<td>2</td>
<td>6.105</td>
<td>3.216</td>
<td>1 v 3</td>
<td>.938</td>
</tr>
<tr>
<td>3</td>
<td>6.403</td>
<td>2.939</td>
<td>1 v 4</td>
<td>.928</td>
</tr>
<tr>
<td>4</td>
<td>6.614</td>
<td>3.143</td>
<td>2 v 3</td>
<td>.945</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>6.342</td>
<td>3.196</td>
<td>2 v 4</td>
<td>.932</td>
</tr>
<tr>
<td>3 v 4</td>
<td></td>
<td></td>
<td>3 v 4</td>
<td>.960</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>6.342</td>
<td></td>
<td>.938</td>
</tr>
<tr>
<td><strong>2. Sit-Ups (No.)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>31.438</td>
<td>13.779</td>
<td>1 v 2</td>
<td>.825</td>
</tr>
<tr>
<td>2</td>
<td>31.807</td>
<td>15.370</td>
<td>1 v 3</td>
<td>.858</td>
</tr>
<tr>
<td>3</td>
<td>34.543</td>
<td>16.025</td>
<td>1 v 4</td>
<td>.757</td>
</tr>
<tr>
<td>4</td>
<td>39.596</td>
<td>21.543</td>
<td>2 v 3</td>
<td>.943</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>34.346</td>
<td>16.703</td>
<td>2 v 4</td>
<td>.829</td>
</tr>
<tr>
<td>3 v 4</td>
<td></td>
<td></td>
<td>3 v 4</td>
<td>.890</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>34.346</td>
<td></td>
<td>.861</td>
</tr>
<tr>
<td><strong>3. Standing Broad Jump (Ins.)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>84.508</td>
<td>9.300</td>
<td>1 v 2</td>
<td>.823</td>
</tr>
<tr>
<td>2</td>
<td>86.543</td>
<td>10.390</td>
<td>1 v 3</td>
<td>.822</td>
</tr>
<tr>
<td>3</td>
<td>87.333</td>
<td>9.474</td>
<td>1 v 4</td>
<td>.820</td>
</tr>
<tr>
<td>4</td>
<td>87.877</td>
<td>9.959</td>
<td>2 v 3</td>
<td>.937</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>86.565</td>
<td>9.785</td>
<td>2 v 4</td>
<td>.941</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 v 4</td>
<td>.951</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>86.565</td>
<td></td>
<td>.899</td>
</tr>
</tbody>
</table>
TABLE I - Continued

<table>
<thead>
<tr>
<th>Variable and Trials</th>
<th>Mean Scores</th>
<th>Standard Deviations</th>
<th>Between Trials</th>
<th>Test-Retest Reliability Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Shuttle Run (Secs.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9.696</td>
<td>.557</td>
<td>1 v 2</td>
<td>.761</td>
</tr>
<tr>
<td>2</td>
<td>9.484</td>
<td>.449</td>
<td>1 v 3</td>
<td>.686</td>
</tr>
<tr>
<td>3</td>
<td>9.464</td>
<td>.469</td>
<td>2 v 3</td>
<td>.806</td>
</tr>
<tr>
<td>4</td>
<td>9.292</td>
<td>.439</td>
<td>2 v 4</td>
<td>.878</td>
</tr>
<tr>
<td>Average</td>
<td>9.484</td>
<td>.481</td>
<td>3 v 4</td>
<td>.754</td>
</tr>
</tbody>
</table>

| 5. 50-Yard Dash (Secs.) |             |                     |                |                                     |
| 1.                   | 6.812       | .437                | 1 v 2          | .797                                |
| a2                   | 6.899       | .428                | 1 v 3          | .804                                |
| 3                   | 6.712       | .427                | 1 v 4          | .762                                |
| 4                   | 6.750       | .436                | 2 v 3          | .770                                |
| Average             | 6.793       | .432                | 2 v 4          | .755                                |
|                      |             |                     | 3 v 4          | .850                                |
|                      |             |                     |                | .792                                |

| 6. Softball Throw (Ft.) |             |                     |                |                                     |
| 1                   | 164.917     | 34.538              | 1 v 2          | .960                                |
| a2                   | 157.631     | 32.717              | 1 v 3          | .931                                |
| 3                   | 165.157     | 35.735              | 1 v 4          | .938                                |
| 4                   | 163.631     | 36.004              | 2 v 3          | .935                                |
| Average             | 162.833     | 34.772              | 2 v 4          | .938                                |
|                      |             |                     | 3 v 4          | .933                                |
|                      |             |                     |                | .940                                |

| 7. 600-Yard Run-Walk (Secs.) |             |                     |                |                                     |
| 1                   | 110.438     | 9.888               | 1 v 2          | .764                                |
| a2                   | 111.561     | 11.946              | 1 v 3          | .659                                |
| 3                   | 107.175     | 11.138              | 1 v 4          | .860                                |
| 4                   | 108.350     | 10.797              | 2 v 3          | .656                                |
| Average             | 108.881     | 10.967              | 2 v 4          | .803                                |
|                      |             |                     | 3 v 4          | .814                                |

\[Trial two scores were adversely affected by wet weather conditions\]
Pull-Ups

For the group tested improvement in mean pull-up performance over four trials was approximately one-half pull-up. The between trials reliability coefficients showed a progressive improvement from .917 to .960. The variability remained relatively constant. It would appear, therefore, from this data that only one or two practice trials are necessary before test results can be used in experimental studies.

Sit-Ups

Mean sit-up performance increased from 31.438 to 39.596 over four trials. This was an increase of approximately eight sit-ups. The standard deviations increased from 13.779 to 21.543. The between trial reliability coefficients varied considerably with no particular pattern evident. The large improvement in mean performance and the increased variance over four trials suggest that four or more trials are required before reliable results can be used in experimental studies. This may not be a suitable item for motor fitness tests since motivation and learning appear to have a marked influence on performance.

Standing Broad Jump

The standing broad jump variable showed an increase in mean performance of 3.369 inches. The variability remained relatively constant and the between trials reliability coefficients increased from .823 to .961. It is obvious from this data that good reliability can only be achieved after three or four practice trials.
Shuttle Run

There was considerable reduction in shuttle run mean performance times and variabilities over four trials and one can only speculate how much more improvement might have resulted from further practice. It is obvious this item needs plenty of practice before the skills involved in changing direction and picking up and placing blocks can be learned adequately.

50-Yard Dash

Trial two scores were obviously affected by wet weather conditions; however, the means and standard deviations of the other trials remained relatively constant. These results indicate that this item requires little learning and few practice trials are necessary before reliable results can be achieved. The event is much simpler than the shuttle run and much less learning is involved.

Softball Throw

Poor weather conditions caused a reduction in the mean and standard deviation of trial two scores. The event was otherwise highly reliable as subjects showed very little improvement in performance over four trials. From these results it is reasonable to assume that trial one scores are sufficiently reliable for experimental purposes.

600-Yard Run-Walk

The small improvement in mean performance (.820 seconds) and the
reasonably large reliability coefficient (.814) between trials three and four indicate stable results and suggest that at least two preliminary practice trials are necessary before recording representative times.

II. STANDARD ERRORS OF MEASUREMENT

Table II contains the standard errors of measurement of the test items as calculated by standard correlation formula method and the analysis of variance technique.

**TABLE II**

AAHPER YOUTH FITNESS TEST STANDARD ERRORS OF MEASUREMENT

<table>
<thead>
<tr>
<th>Variables</th>
<th>Units</th>
<th>Correlation Formula Method</th>
<th>Analysis of Variance Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Pull-Ups</td>
<td>No.</td>
<td>0.794</td>
<td>0.834</td>
</tr>
<tr>
<td>2 Sit-Ups</td>
<td>No.</td>
<td>6.250</td>
<td>6.934</td>
</tr>
<tr>
<td>3 Standing Broad Jump</td>
<td>Ins.</td>
<td>3.124</td>
<td>3.353</td>
</tr>
<tr>
<td>4 Shuttle Run</td>
<td>Secs.</td>
<td>0.227</td>
<td>0.239</td>
</tr>
<tr>
<td>5 50-Yard Dash</td>
<td>Secs.</td>
<td>0.194</td>
<td>0.190</td>
</tr>
<tr>
<td>6 Softball Throw</td>
<td>Ft.</td>
<td>9.100</td>
<td>9.170</td>
</tr>
<tr>
<td>7 600-Yard Run-Walk</td>
<td>Secs.</td>
<td>5.000</td>
<td>5.660</td>
</tr>
</tbody>
</table>

*Standard errors of measurement were calculated with trial two data eliminated.
The standard error of measurement indicates the amount of variation one could expect in a person's score, if he were examined repeatedly with the same test. When interpreting a person's score it is advisable to think in terms of the standard error of measurement and to avoid making any definite conclusions concerning that score. It is important to understand that scores obtained for an individual are not true scores but only estimates of the individual's true score. The standard error of measurement gives us a clear idea of how accurate our measurements are and how much confidence we can place on getting a reasonably accurate measurement.

In the present study the standard errors of measurement calculated by the standard correlation formula method compared closely with those calculated by the analysis of variance technique. The differences between errors of measurement in Table II are probably the result of rounding errors. A choice between the two methods of determining the standard error of measurement will depend to a large extent upon the nature of the problem under consideration.

III. ANALYSIS OF VARIANCE

Each variable was treated by analysis of variance to answer the following questions:

1. Is the practice effect significant?

2. Does the test item measure with an accuracy sufficient to distinguish between the subjects tested?

3. Are subject differences significantly greater than practice differences?
Table III contains an analysis of variance for each variable. Three F ratios have been calculated for each variable ($F_1$, $F_2$, $F_3$). $F_1$ refers to question one, $F_2$ refers to question two, and $F_3$ refers to question three.

**TABLE III**

**AAHPER YOUTH FITNESS TEST ANALYSIS**

**OF VARIANCE DATA**

<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square (Variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Pull-Ups</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Trials (Practice Effect)</td>
<td>3</td>
<td>8.15789</td>
<td>2.71929</td>
</tr>
<tr>
<td>Between Subjects (Individuals)</td>
<td>56</td>
<td>2172.31000</td>
<td>38.79130</td>
</tr>
<tr>
<td>Error (Residual)</td>
<td>168</td>
<td>116.84200</td>
<td>0.69549</td>
</tr>
<tr>
<td>Total</td>
<td>227</td>
<td>2297.3100</td>
<td></td>
</tr>
<tr>
<td>$F_1 = \frac{V. \text{ Trials}}{V. \text{ Error}} = 3.91^a$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_2 = \frac{V. \text{ Subjects}}{V. \text{ Error}} = 55.77^c$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_3 = \frac{V. \text{ Subjects}}{V. \text{ Trials}} = 14.26^d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2. Sit-Ups</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Trials (Practice Effect)</td>
<td>3</td>
<td>2605.10</td>
<td>868.36</td>
</tr>
<tr>
<td>Between Subjects (Individuals)</td>
<td>56</td>
<td>53988.60</td>
<td>964.08</td>
</tr>
<tr>
<td>Error (Residual)</td>
<td>168</td>
<td>8077.90</td>
<td>48.08</td>
</tr>
<tr>
<td>Total</td>
<td>227</td>
<td>64671.60</td>
<td></td>
</tr>
<tr>
<td>$F_1 = \frac{V. \text{ Trials}}{V. \text{ Error}} = 18.06^a$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_2 = \frac{V. \text{ Subjects}}{V. \text{ Error}} = 20.05^c$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_3 = \frac{V. \text{ Subjects}}{V. \text{ Trials}} = 1.11^d$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**TABLE III - Continued**

<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square (Variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing Broad Jump</td>
<td>3</td>
<td>372.820</td>
<td>124.2730</td>
</tr>
<tr>
<td>Between Trials (Practice Effect)</td>
<td>3</td>
<td>19561.200</td>
<td>349.3080</td>
</tr>
<tr>
<td>Between Subjects (Individuals)</td>
<td>56</td>
<td>1887.930</td>
<td>11.2376</td>
</tr>
<tr>
<td>Error (Residual)</td>
<td>168</td>
<td>1887.930</td>
<td>11.2376</td>
</tr>
<tr>
<td>Total</td>
<td>227</td>
<td>21822.000</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
F_1 &= \frac{V. \text{ Trials}}{V. \text{ Error}} = 11.05^a \\
F_2 &= \frac{V. \text{ Subjects}}{V. \text{ Error}} = 31.08^c \\
F_3 &= \frac{V. \text{ Subjects}}{V. \text{ Trials}} = 2.81^d
\end{align*}
\]

**4. Shuttle Run**

<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square (Variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shuttle Run</td>
<td>3</td>
<td>4.67416</td>
<td>1.5580500</td>
</tr>
<tr>
<td>Between Trials (Practice Effect)</td>
<td>3</td>
<td>42.31370</td>
<td>.7556030</td>
</tr>
<tr>
<td>Between Subjects (Individuals)</td>
<td>56</td>
<td>9.60834</td>
<td>.0571925</td>
</tr>
<tr>
<td>Error (Residual)</td>
<td>168</td>
<td>56.59620</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>227</td>
<td>56.59620</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
F_1 &= \frac{V. \text{ Trials}}{V. \text{ Error}} = 27.24^a \\
F_2 &= \frac{V. \text{ Subjects}}{V. \text{ Error}} = 13.21^c \\
F_3 &= \frac{V. \text{ Subjects}}{V. \text{ Trials}} = .48^d
\end{align*}
\]

**5. 50-Yard Dash**

<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square (Variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-Yard Dash</td>
<td>2</td>
<td>0.29</td>
<td>0.150</td>
</tr>
<tr>
<td>Between Trials (Practice Effect)</td>
<td>2</td>
<td>27.54</td>
<td>0.490</td>
</tr>
<tr>
<td>Between Subjects (Individuals)</td>
<td>56</td>
<td>4.11</td>
<td>0.036</td>
</tr>
<tr>
<td>Error (Residual)</td>
<td>112</td>
<td>31.94</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
<td>31.94</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
F_1 &= \frac{V. \text{ Trials}}{V. \text{ Error}} = 4.16^a \\
F_2 &= \frac{V. \text{ Subjects}}{V. \text{ Error}} = 13.61^c \\
F_3 &= \frac{V. \text{ Subjects}}{V. \text{ Trials}} = 3.27^d
\end{align*}
\]
<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square (Variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>6. Softball Throw</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Trials (Practice Effect)</td>
<td>2</td>
<td>117.93</td>
<td>58.965</td>
</tr>
<tr>
<td>Between Subjects (Individuals)</td>
<td>56</td>
<td>202402.28</td>
<td>3614.330</td>
</tr>
<tr>
<td>Error (Residual)</td>
<td>112</td>
<td>9420.74</td>
<td>94.110</td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
<td>211940.95</td>
<td></td>
</tr>
<tr>
<td>$F_1 = \frac{V. Trials}{V. Error} = 0.701^b$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_2 = \frac{V. Subjects}{V. Error} = 42.97^c$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_3 = \frac{V. Subjects}{V. Trials} = 61.29^d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>7. 600-Yard Run-Walk</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Trials (Practice Effect)</td>
<td>2</td>
<td>532.71</td>
<td>266.35</td>
</tr>
<tr>
<td>Between Subjects (Individuals)</td>
<td>56</td>
<td>16115.98</td>
<td>287.79</td>
</tr>
<tr>
<td>Error (Residual)</td>
<td>112</td>
<td>3589.29</td>
<td>32.05</td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
<td>20237.98</td>
<td></td>
</tr>
<tr>
<td>$F_1 = \frac{V. Trials}{V. Error} = 8.31^a$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_2 = \frac{V. Subjects}{V. Error} = 8.97^c$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_3 = \frac{V. Subjects}{V. Trials} = 1.09^d$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**a** Significant at the 0.05 level of confidence.

**b** Not significant at the 0.05 level of confidence.

**c** Significant at the 0.01 level of confidence.

**d** Not significant at the 0.01 level of confidence.

**NOTE:** F ratios were calculated with trial two data eliminated.
If a true hypothesis that there was no practice effect over four trials was rejected, there is an error of the first kind (alpha error). If a false hypothesis that there was no practice effect over four trials was accepted, there would be an error of the second kind (beta error). Logic and experience suggest that there would be practice effect over four trials in most motor fitness tests performed by untrained subjects. It is considered conservative procedure in motor fitness testing to train subjects in test performance before recording results. One would be reluctant to make the error that there was no practice effect over four trials if in truth this effect did occur. One would not be so reluctant to make the error that there was practice effect over four trials if in truth there was no real practice effect. For the above reasons a level of confidence of 0.05 rather than 0.01 was set for F tests of practice effect.

Analysis of variance results indicate a significant practice effect for all test items except the softball throw at the 0.05 level of confidence. The value of F for the indoor testing items (pull-ups, sit-ups, standing broad jump and shuttle run) from the F Table with 3 and 168 degrees of freedom at 5 percent level of confidence is 2.66. The value of F for the outdoor testing items (50-yard dash, softball throw and 600-yard run-walk) from the F Table with 2 and 112 degrees of freedom at the 5 percent level of confidence is 3.08. The values of $F_1$ (practice effect) obtained for each variable were pull-ups 3.91, sit-ups 18.06, standing broad jump 11.05, shuttle run 27.24, 50-Yard dash 4.16, softball throw 0.701, and 600-yard run-walk
As the value of $F_1 (0.701)$ obtained for the softball throw is smaller than the value of $F (3.08)$ from the F Table at the 5 percent level of confidence with 2 and 112 degrees of freedom, it can be concluded that the practice effect is not significant. Since the values of $F_1$ obtained for the indoor test items are larger than the value of $F (2.66)$ at the 5 percent level of confidence with 3 and 168 degrees of freedom and since the values of $F_1$ obtained for the 50-yard dash and the 600-yard run-walk are larger than the value of $F (3.08)$ at the 5 percent level of confidence with 2 and 112 degrees of freedom, it can be concluded that the practice effect is significant for all test items except the softball throw.

The existence of practice effect is of considerable importance and should be considered in all research studies. The significant practice effect in this study indicates that much learning has taken place and that all test items with the possible exception of the softball throw should be practiced before stable results can be recorded and interpreted.

Differences Between Subjects ($F_2$)

A level of confidence of 0.01 was set for the between subjects F ratios. A motor performance test item should be able to show consistent discrimination between individual differences in ability from trial to trial and it should do this despite random fluctuation in scores due to errors of measurement. The high level of confidence of 0.01 was chosen so that there should be as little doubt as possible about the accuracy of the test item in distinguishing between the abilities of the subjects tested.
Analysis of variance results indicate that the test items measure with an accuracy sufficient to distinguish between the subjects tested. The value of F for the indoor test items (pull-ups, sit-ups, standing broad jump and shuttle run) from the F Table with 56 and 168 degrees of freedom at the 1 percent level of confidence is 1.62. The value of F for the outdoor test items (50-yard dash, softball throw, and 600-yard run-walk) from the F Table with 56 and 112 degrees of freedom at the 1 percent level of confidence is 1.69. The values of \( F_2 \) (subjects) obtained for each variable from the study were pull-ups 55.77, sit-ups 20.05, standing broad jump 31.08, shuttle run 13.21, 50-yard dash 13.61, softball throw 42.97, 600-yard run-walk 8.97. Since the values of \( F_2 \) obtained for the indoor test items are larger than the value of F (1.62) at the 1 percent level of confidence with 56 and 168 degrees of freedom and since the values of \( F_2 \) obtained for the outdoor items are larger than the value of F (1.69) at the 1 percent level of confidence with 56 and 112 degrees of freedom, it can be concluded that for all items the between subjects mean square (variance) is significantly larger than the error mean square (variance).

Since the subject mean square for each test item was significantly larger than the error mean square, it can be assumed that the differences between subject scores were so large that they could not be due entirely to errors of measurement. Thus, it can be further concluded that the test items measure with an accuracy sufficient to distinguish between the abilities of the subjects tested.
Subject Differences \( \text{Versus Practice Differences (F_g)} \)

The 0.01 level of confidence was chosen for the F ratio, subject variance over practice (between trials) variance, in order to guard against the possibility of rejecting a true hypothesis that practice effect was severe. Normally subject variance is much larger than practice variance in repeated trials of motor performance tests. A non-significant F ratio would indicate that this was not so and that the practice effect is severe. If practice effect is severe it would be desirable to show this as a warning to test administrators that adequate preliminary practice is needed before performances could be considered sufficiently stable for use.

Analysis of variance results indicate that subjects differences are not significantly greater than practice differences. The value of F for the indoor test items (pull-ups, sit-ups, standing broad jump and shuttle run) from the F Table with 56 and 3 degrees of freedom at the 1 percent level of confidence is 26.32. The value of F for the outdoor items (50-yard dash, softball throw and 600-yard run-walk) from the F Table with 56 and 2 degrees of freedom at the 1 percent level of confidence is 99.49. The values of F_g (subject variance over practice variance) obtained for each variable from the study were pull-ups 14.26, sit-ups 1.11, standing broad jump 2.81, shuttle run .48, 50-yard dash 3.27, softball throw 61.29, and 600-yard run-walk 1.08. Since the values of F_g obtained for the indoor test items are smaller than the value of F (26.32) at the 1 percent level of confidence with 56 and 3 degrees of freedom and since the values of F_g obtained for the outdoor items are smaller than the value of F (99.49) at the 1 percent level of confidence
with 56 and 2 degrees of freedom, it can be concluded that for all test items subject differences are not significantly larger than practice differences (differences between trials).

As the between subjects mean square (variance) is not significantly larger than the between trials mean square (variance), and since it usually is, the implication is that the practice effect as a result of learning and other factors must have been severe. This fact is verified by the significant between trials (practice effect) F ratios for all items except the softball throw.

The non significant $F_3$ ratio (between subject variance over between trials variance) for the softball throw indicates some practice has occurred. However, since the size of this ratio closely approaches the one percent level of confidence, it can be regarded as almost significant. Hence, the amount of practice that has occurred is very small and certainly not large enough to show significant differences between trial means.
CHAPTER VI

SUMMARY AND CONCLUSIONS

The object of this study was to determine by a reliability analysis if the American Association for Health, Physical Education and Recreation Youth Fitness Test items are accurate and reliable measuring instruments.

This study was made possible through the use of fifty-seven untrained male first and second year students in the Required Physical Education Programme at the University of British Columbia who were tested once a week for four consecutive weeks with the AAHPER Test.

The data from each test item were analyzed in order to obtain (a) means and standard deviations on each of four trials, (b) between trials correlation coefficients and an average reliability coefficient, (c) standard errors of measurement computed by the standard correlation formula method and the analysis of variance technique, and (d) three F ratios (analysis of variance).

The findings of this study based on the reliability analysis of the data collected may be summarized as follows:

1. Test items with average test-retest reliability coefficient over four trials of .85 or better were pull-ups (.938), sit-ups (.861), standing broad jump (.899), and softball throw (.940). Items with the poorest average test-retest reliability coefficients over four trials were shuttle run (.776), 50-yard dash (.792), and the 600-yard run-walk (.759).
2. The best between trial reliability coefficient for each variable was pull-ups \( (T_3 - T_4 = .960) \), sit-ups \( (T_2 - T_5 = .943) \), standing broad jump \( (T_5 - T_4 = .951) \), shuttle run \( (T_2 - T_4 = .878) \), 50-yard dash \( (T_3 - T_4 = .850) \), softball throw \( (T_1 - T_2 = .960) \), and 600-yard run-walk \( (T_1 - T_4 = .860) \).

3. The standard errors of measurement of the test items computed by the standard correlation formula method compared closely with those computed by the analysis of variance technique.

4. The standard errors of measurement computed by the standard correlation formula method and the analysis of variance technique were pull-ups (correlation formula method 0.794 and analysis of variance technique 0.834), sit-ups (6.250 and 6.934), standing broad jump (3.124 and 3.353 inches), shuttle run (0.227 and 0.239 seconds), 50-yard dash (0.194 and 0.190 seconds), softball throw (9.100 and 9.170 feet), and 600-yard run-walk (5.000 and 5.660 seconds).

5. Analysis of variance results showed a significant practice effect over four trials for all test items except the softball throw.

6. Analysis of variance results showed that the AAHPER Test items measure with an accuracy sufficient to distinguish between the subjects tested.

7. Analysis of variance results showed that for each test item subject differences are not significantly larger than practice differences. Since subject differences are usually much larger than practice differences, it can be concluded that practice effect in this study must have been severe.
In accordance with the above findings the following recommendations are made to anyone contemplating administering the AAHPER Test.

1. The pull-ups and softball throw variables appear to be highly reliable measuring instruments. Thus for these items it seems reasonable to accept first trial scores as sufficiently accurate for both survey and experimental purposes.

2. The standing broad jump, 50-yard dash and 600-yard run-walk items are apparently not as reliable as the pull-ups and softball throw items and several practice trials should be given before results can be accepted as reasonably accurate estimates of true individual performance.

3. The sit-ups and shuttle run test items are apparently the least reliable items of the AAHPER Test. These items seem to require at least four, perhaps more, preliminary practice trials before a satisfactory level of reliability can be attained.
BIBLIOGRAPHY

A. BOOKS


B. PERIODICALS


C. PUBLICATIONS OF LEARNED SOCIETIES


Jackson, Robert W. B. Application of the Analysis of Variance and Covariance Method to Education Problems. University of Toronto, Toronto, Department of Educational Research, Bulletin No. 11, 1940.

Jackson, Robert W. B., and George A. Ferguson. Studies on the Reliability of Tests, University of Toronto, Toronto, Department of Educational Research, Bulletin No. 12, 1941.

D. UNPUBLISHED MATERIALS


APPENDIX
APPENDIX A

AAHPER RESEARCH PROGRAMME

THE UNIVERSITY OF BRITISH COLUMBIA

School of Physical Education and Recreation

Required Physical Education Program, P. E. 64

March 5, 1962.

During the final weeks of this term the School of Physical Education and Recreation will be administering physical fitness tests several times to students enrolled in P. E. 64. One test is that of the American Association for Health, Physical Education and Recreation which has been widely used in the schools of the U.S. and other countries in recent years. It has also been used in universities and scoring scales for university students are available. The items of this test are pull-ups on horizontal bar, sit-ups (up to 100), standing broad jump, shuttle run (4 x 30 feet) - all done indoors, and 50-yard sprint, softball throw, 600-yard run - all done outdoors. This test will be administered four times. The indoor part will be administered once during the first part of each week and the outdoor part will be administered once in the second part of each week.

You have been scheduled to attend testing sessions twice a week until the end of term. Testing begins Monday 12th March. It is absolutely necessary that you do not miss a session unless absence is completely unavoidable. Absentees must be explained in writing or an excuse will not
be recorded. If a scheduled session is missed for a very special reason, it will be possible to "make it up" in the same week.

"Make-up" times are as follows:

Indoor part - Wednesdays, from 12:30 to 1:30 or 4:30 to 6:30 p.m.

Outdoor part - Fridays, from 4:30 to 6:30 p.m.

If you are scheduled to attend Saturday but will have to miss, you must attend the day before - i.e. on Friday at 4:30.

Be certain to change into shorts, shirt and tennis or basketball shoes for every session. It is advisable also to have a sweat suit for the outdoor sessions. If you forget to bring equipment, attendance is still necessary. On arrival at the War Memorial Gymnasium on testing days, consult the blackboard in the foyer for directions about what to do or where to go when changed. Any cancellations will be posted on the blackboard.

Unfortunately, the weather has not permitted testing to be carried out earlier in the term. Your consistent attendance and co-operation during the next few weeks will be greatly appreciated and will ensure that your physical education requirement will be completed satisfactorily.

You have been scheduled to attend testing sessions at the War Memorial Gymnasium during the weeks beginning Monday March 12th, 19th, 26th, and April 2nd, on
Mondays at ________________ Fridays at ________________

Wednesdays at ________________ Saturdays at ________________

Stanley R. Brown,
Assistant Professor.
APPENDIX B

HISTORY AND USE OF THE AAHPER TEST

THE UNIVERSITY OF BRITISH COLUMBIA
School of Physical Education and Recreation

The AAHPER Physical (Motor) Fitness Test

The AAHPER Test is being widely used in schools and universities of the North American Continent and Overseas. It came into existence to satisfy the need for a suitable test to be used in conjunction with the "fitness movement" which started with the support of President Eisenhower and has continued with the enthusiastic promotion of President Kennedy. Recently the Government of Canada voted an annual grant of five million dollars to promote fitness and sport among the youth and adults of this country.

By "motor fitness" we mean the capacity for efficient performance in the basic requirements of running, jumping, dodging, falling, climbing, swimming, lifting weights (including one's own body), carrying loads and enduring under sustained effort in a variety of situations. The AAHPER Test is designed to measure a number of these attributes.

The project in which you are now a participant is designed to find out some important facts about the AAHPER Test in the event it is put into general use in this University. It is most important that you should do your best on all occasions. Only in this way will it be possible for you to find out what your true capabilities are.
Percentile scoring scales for college men have been published recently. When the tests are over you will be given a record of your results and percentile ratings.

Full attendance and genuine effort over the four weeks of the project will be greatly appreciated and will ensure that you will complete your physical education requirement for the year.
APPENDIX C

AAHPER TEST ADMINISTRATION

INSTRUCTIONS

**Pull-Ups:** This test measures the strength and endurance of the arm and shoulder muscles. Use the overhand grip - palms away from the body. From a full hang, with arms and legs fully extended, raise the body until the chin can be placed over the bar. Lower to a full hang. Repeat continuously without pause as many times as possible. When tired do not drop from the bar from the "chinning" position. Lower to a full hang and attempt to pull up again. If unsuccessful lower to a full hang and then drop from the bar.

**Cautions:** Swinging the body, raising or kicking the knees or legs is not permitted.

**Sit-Ups:** This test measures the strength and endurance of the hip flexor and abdominal muscles.

**Position:** Lie on the back, legs extended, feet about two feet apart, hands behind the neck, fingers interlocked, elbows retracted and touching the floor.

A partner holds the ankles down, the heels being on the floor at all times.

**Execution:** Sit up, turning the trunk to the left and touch the right elbow to the left knee. Return to the starting position. Sit up, turning the trunk to the right and touch the left elbow to the right knee. Return to the starting position.
Repeat the exercise continuously without pause alternating to the left and right.

**Cautions:** Fingers must remain in contact behind the neck throughout. The knees must be on the floor during the sit-ups but may be slightly bent when touching elbow to knee. When returning to the starting position, the elbows must be flat on the floor before sitting up again.

**Partner:** Do not count the sit-up if the finger tips do not remain in contact behind the head throughout or if the knees are bent when the subject begins to sit up.

Count the number of correct sit-ups. Stop the person you are holding at 100 sit-ups or whenever he is unable to continue without resting longer than two seconds in the starting or sit-up position.

**Standing Broad Jump:** This test measures ability to move the body "explosively". It reflects speed and strength in propelling the bodyweight.

**Position:** Stand with the feet slightly apart and toes just behind the take-off line.

**Execution:** Preparatory to jumping swing the arms backward and bend the knees. Jump as far forward as possible by simultaneously extending the knees and swinging the arms forward.

Three trials are allowed. Distance is measured from take-off line to heel or other part of the body that touches the floor nearest the take-off line.

**Shuttle Run:** This test measures the ability to move quickly and change direction i.e. speed plus agility.
Two lines are drawn 30 feet apart. Two blocks of wood are placed just behind one line and the start takes place behind the other line.

**Position:** Take up a standing start position with the front foot just behind the starting line.

**Execution:** On the signal "Ready? Go!" run to the opposite line, pick up one block, return and place it just over the starting line. Return to pick up the other block and race over the starting line, block in hand.

Two trials are allowed, with a rest in between.

**Caution:** Perform the test at full speed. When finishing do not slow down but run through the finish as fast as possible. Do not place the second block down but keep it in your hand.

Throwing or dropping the first block is not allowed.

**50-Yard Dash:** This test measures speed in running.

**Position:** Take up a standing start position with the front foot just behind the starting line.

**Execution:** The starter will use the commands "Are you ready?" and "Go!" The latter will be accompanied by a downward sweep of the arm to give the timekeeper a visual signal.

Run as fast as you can maintaining full speed past the finishing line.

Two trials are allowed, with a rest in between.

**Softball Throw for Distance:** This test measures co-ordination as well as arm strength and speed.

**Directions:** The test consists of three consecutive throws for distance. The throwing area is confined within two parallel lines, six feet apart, at
right angles to the direction of the throw. Stepping on or over the lines during a throw or on following through constitutes a foul.

Only overhand throws are permitted.

600-Yard Run: This test measures running endurance and reflects will-power, muscular and circulatory - respiratory fitness and pace judgement.

Directions: Before this event you will be paired with a partner who will hold your score card during the event and will place himself opposite the finish line, ready to record your time as it is called out by the timer. When you have finished your run, return to the finish line (a) ready to perform the same function for your partner if he has not already run or (b) to collect your score card.

Execution: At the signal "Ready?, Go!" begin running from a standing start and maintain the best pace possible throughout the event to finish in the best possible time.
APPENDIX D

AAHPER TEST SCORE CARD

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty/Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School/City</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pull-Ups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sit-Ups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standing Broad Jump</td>
<td>1. 2. 3.</td>
<td>1. 2. 3.</td>
<td>1. 2. 3.</td>
<td>1. 2. 3.</td>
</tr>
<tr>
<td>Shuttle Run</td>
<td>1. 2.</td>
<td>1. 2.</td>
<td>1. 2.</td>
<td>1. 2.</td>
</tr>
<tr>
<td>50-Yard Dash</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Softball Throw</td>
<td>1. 2. 3.</td>
<td>1. 2. 3.</td>
<td>1. 2. 3.</td>
<td>1. 2. 3.</td>
</tr>
<tr>
<td>600-Yard Run</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>