

## **INFORMATION TO USERS**

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

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

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

**Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.**

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

**Bell & Howell Information and Learning  
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA  
800-521-0600**

**UMI<sup>®</sup>**



Examining the Seasonal Variation and Reliability of Health-  
Related Fitness Scores in Children using a Multivariate  
Model

J.P. Barfield

A dissertation presented to the Graduate Faculty of Middle  
Tennessee State University in partial fulfillment of the  
requirements for the Doctor of Arts degree in Physical  
Education in the Department of Health, Physical Education,  
Recreation, and Safety.

August, 2000

UMI Number: 9978693

**UMI<sup>®</sup>**

---

UMI Microform 9978693

Copyright 2000 by Bell & Howell Information and Learning Company.

All rights reserved. This microform edition is protected against  
unauthorized copying under Title 17, United States Code.

---

Bell & Howell Information and Learning Company  
300 North Zeeb Road  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

Examining the Seasonal Variation and Reliability of Health-  
Related Fitness Scores in Children using a Multivariate  
Model

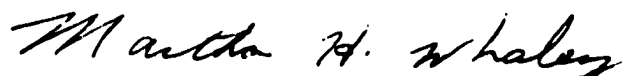
APPROVED:


Graduate Committee:

  
Major Professor

  
Committee Member

  
Committee Member

  
Head of the Department of Health, Physical Education,  
Recreation, and Safety

  
Dean of the Graduate College

## ABSTRACT

### Examining the Relationships between Fall and Spring Health-Related Fitness Scores in Elementary School-Aged Children using a Multivariate Model

The FITNESSGRAM is the latest test battery associated with the American Alliance for Health, Physical Education, Recreation, and Dance (AAHPERD) that is purported to measure the physical fitness of children. To ensure that the battery is an appropriate assessment tool, validity and reliability evidence must be established. To this point, no battery reliability evidence of the FITNESSGRAM test has been reported. Additionally, it is important to document children's fitness changes during elementary school years from a multivariate perspective (i.e., fitness as one trait) and not strictly from a univariate perspective (i.e., separate interpretations for the one-mile run, pull-up, etc). The purpose of the study was to: 1) determine the optimal reliability of the FITNESSGRAM battery among children in grades 3 through 6; 2) evaluate the inter-individual changes in health-related fitness scores among children in grades 3 through 6 across a school year; and 3) evaluate intra-individual changes in health-related fitness

scores among children in grades 3 through 6 across a school year.

The FITNESSGRAM test battery appears to be a reliable instrument to measure health-related fitness among children in grades 3 through 6. Children's health-related fitness, as a multivariate construct, is only moderately stable during the academic school year, with maturational changes impacting the stability of scores among children in grades 3 and 4 and motivational factors hampering stability among children in grade 6. Children in grade 5, however, seem to have the most stable health-related fitness scores over time. Finally, although children are classified into the same healthy/unhealthy categories from fall to spring fairly consistently, evidence exists that improvement in health-related fitness may not result from physical education classes only twice per week, especially in terms of body fatness.

## ACKNOWLEDGEMENTS

To Mom and Dad - all my love. My gratitude will never equal all you have done for me.



## TABLE OF CONTENTS

	Page
List of Tables .....	vi
List of Appendices .....	vii
I. Introduction .....	1
Statement of the Problem .....	4
Purpose .....	5
Research Questions .....	5
Assumptions .....	6
Delimitations .....	6
Definition of Terms .....	6
II. Review of the Literature .....	10
History of Fitness Testing among School Children .....	11
Objectives of Fitness Testing in Schools ....	16
The Tracking of Health-related Fitness Scores .....	19
Reliability Theory .....	21
III. Method .....	31
Subjects .....	31
Instruments .....	32
Procedures .....	33
Analyses .....	40

IV. Results .....	41
Optimal Reliability .....	42
Inter-individual Changes .....	44
Intra-individual Changes .....	46
IV. Discussion .....	50
Optimal Reliability .....	50
Inter-individual Changes .....	51
Intra-individual Changes .....	54
Comparison to National Probability Studies ..	56
Suggestions for Future Research .....	58
Summary .....	59
References .....	94

## List of Tables

Table	Page
1. Means and Standard Deviations for the Entire Sample .....	41
2. Optimal Reliability Coefficients .....	43
3. Shared Variance between Fall and Spring Test Battery Composite Scores for the Entire Sample .....	44
4. Shared Variance between Fall and Spring Test Battery Composite Scores for Boys	45
5. Shared Variance between Fall and Spring Test Battery Composite Scores for Girls	45
6. Percent Agreement of Combined Sample on Criterion Standards .....	47
7. Percent Agreement of Boys on Criterion Standards .....	48
8. Percent Agreement of Girls on Criterion Standards .....	48

## List of Appendices

Appendix	Page
A. Test-retest Reliability Estimates of Physical Tests Included in the FITNESSGRAM .....	60
B. Descriptive Statistics on the Individual Test Items .....	68
C. Frequency of Healthy Classification on FITNESSGRAM Standards .....	76
D. Approval Letters for School Personnel .....	83
E. Institutional Review Board Approval .....	93

## CHAPTER 1

### Introduction

Kraus and Hirschland (1954) first documented muscle and flexibility deficiencies among American children compared to their European peers. Although not a comprehensive evaluation of total physical fitness, the Kraus-Weber test used by these authors sparked increased attention on fitness levels of American children. Upon these findings, President Eisenhower initiated a task force to investigate fitness shortcomings and promote physical fitness among American school children (American Association for Health, Physical Education, & Recreation, 1958; 1965). Determining the change in American youth fitness levels since that time has been a complex process, partly because assessment techniques have varied (Morrow, 1992).

In an attempt to determine whether or not American children's fitness levels have changed since Kraus and Hirschland's findings, an extensive review and commentary of health-related physical fitness was published in *Research Quarterly* (1992) that yielded equivocal opinions. Some authors indicated that health-related physical fitness levels among children have not declined over the past four decades, with the exception of body fatness (Blair, 1992;

Corbin & Pangrazi, 1992). However, others suggested that fitness levels among children have declined (Kuntzleman & Reiff, 1992; Updyke, 1992).

One of the difficulties in determining fitness progress has been the changing definition of physical fitness (Corbin & Pangrazi, 1992). Fitness test items, as well as standards used for evaluation, have changed in response to the evolving definition. Hence, various test batteries have included different test items as well as scoring systems to measure fitness, making comparisons across time (or across test batteries) difficult (Updyke, 1992). Further, the last comprehensive fitness assessment of elementary school children was completed in 1987 (National Child and Youth Fitness II Study).

The Youth Fitness Test (AAHPER, 1958) was the first battery used to measure physical fitness following Kraus and Hirschland's (1954) report. During that decade, fitness was defined as possessing adequate motor ability (Pate, 1983). The battery originated in 1958 and included normative standards of performance based on sample data (n=8,500) of children in grades 5 through 12. Minor adjustments were made to the original Youth Fitness Test in 1965 and 1976 (AAHPER, 1965; AAHPER, 1976). Normative standards were published in both revised batteries with the

intention of assessing improvement in motor skills among children from 1965 to 1976.

After the 1976 battery was developed, the philosophy of physical fitness changed from a motor skill emphasis (e.g., agility, balance) to an emphasis on health promotion and disease prevention (Pate, 1983; 1994). Corresponding to the change in definition, the Youth Fitness Test was replaced by the Health-Related Physical Fitness Test in 1980 (AAHPERD, 1980). The Health-Related Physical Fitness Test battery delineated three components of fitness: cardiorespiratory function, body composition, and abdominal and low back-hamstring musculoskeletal function. Test items reflected the change to a health-related emphasis and new normative performance standards were published for the alternative items. The Prudential FITNESSGRAM (Cooper Institute of Aerobics Research [CIAR], 1992; 1999) has been the latest physical fitness test battery promoted by AAHPERD. Unlike the previous batteries, the FITNESSGRAM includes criterion-referenced standards for performance rather than norm-referenced standards. These Criterion-referenced standards classify individuals as either healthy or unhealthy on a particular test item (i.e., trait).

If fitness tests continue to be used to assess physical fitness, reliability and validity evidence will be

needed for current tests (Seefeldt & Vogel, 1989; Safrit, 1990). For a test to be considered valid, it must have evidence of reliability (Safrit, 1976; Safrit & Wood, 1995). Typically, test battery reliability has been determined by estimating the reliability of the individual test items and applying these results to the test battery itself. Wood and Safrit (1984; 1987) suggested that a canonical correlation analysis (CCA) be used to estimate the reliability of a set of physical tests (or test battery) and may be much more appropriate for assessing battery test-retest reliability than the aforementioned method. Safrit and Wood (1987) conducted a CCA on the Health-Related Physical Fitness Test (AAHPERD, 1980) and concluded that the battery was highly reliable among children in grades 6 through 8. No such reliability evidence has been estimated for the FITNESSGRAM test battery, which is the fitness battery currently promoted by AAHPERD.

#### **Statement of the Problem**

The FITNESSGRAM is the latest test battery associated with AAHPERD that is purported to measure the physical fitness of children. To ensure that the battery is an appropriate assessment tool, validity and reliability evidence must be established. To this point, no battery



reliability evidence of the FITNESSGRAM test has been reported. Additionally, it is important to document children's fitness changes during elementary school years from a multivariate perspective (i.e., fitness as one trait) and not strictly from a univariate perspective (i.e., separate interpretations for the one-mile run, pull-up, etc).

### **Purpose of the Study**

The purpose of the study was to:

- 1) determine the optimal reliability of the FITNESSGRAM battery among children in grades 3 through 6;
- 2) evaluate the inter-individual changes in health-related fitness scores among children in grades 3 through 6; and
- 3) evaluate intra-individual changes in health-related fitness scores among children in grades 3 through 6.

### **Research Questions**

Does the FITNESSGRAM reliably measure physical fitness levels among children?

Do health-related fitness scores change consistently during a school year among children in the same grade?

Do children's health-related fitness classifications, as categorized by the FITNESSGRAM, change across a school year?

**Assumptions**

1. Children involved in testing have practiced the test items sufficiently; therefore, no carry-over effect will result from performing the tests on multiple occasions.

2. Children will give a maximal effort on all trials of the test items.

3. The sample size is sufficient for each grade level to calculate a stable canonical correlation coefficient.

**Delimitations**

1. Children were recruited via convenience sampling from Rutherford county.

2. Subjects were recruited from general education classrooms and data from individuals within special education classrooms were not included in the analysis.

3. The measurement of physical fitness was delimited to the test items and corresponding protocols published in the FITNESSGRAM.

4. Only children in grades 3 through 6 were selected to participate.

**Definition of Terms****Health-related (Physical) Fitness**

The current purpose of health-related fitness is to enhance health and prevent disease (Corbin & Pangrazi,

1992). Health-related physical fitness is commonly defined as a "state characterized by: (1) an ability to perform daily activities with vigor, and (2) traits and capacities that are associated with low risk of premature development of the hypokinetic diseases," (Pate & Shephard, 1989, p. 4). The aforementioned state can be described as possessing minimum levels of health-related fitness on traits that yield good health and disease prevention (e.g., flexibility and muscular strength).

Physical fitness was defined operationally as test item scores for aerobic capacity, body composition, abdominal strength, trunk extension strength/flexibility, upper body strength, and hamstring flexibility. The test scores were measured according to the FITNESSGRAM (CIAR, 1999) battery protocol.

Aerobic capacity. Aerobic capacity is the ability to perform prolonged periods of exercise (American College of Sports Medicine, 1995). Increasing aerobic capacity increases work capacity and decreases risk of coronary heart disease (Pate & Shephard, 1989). Aerobic capacity was defined operationally as the time needed to complete the one-mile run.

Body composition. Body composition is an estimation of an individual's fat mass and lean mass percentage (CIAR,

1999). A high fat body composition is an indication of obesity and associated complications. Body composition was defined operationally as the Body Mass Index ratio.

Abdominal strength. Abdominal strength ensures proper posture and alignment, thereby maintaining efficient low back function (CIAR, 1999). Abdominal strength was defined operationally as the total amount of curl-ups performed to a cadence of one per every three seconds.

Trunk extension/flexibility. Trunk extensor muscles and joint flexibility compliment abdominal muscles in ensuring correct vertebral alignment and sufficient low back function (CIAR, 1999). Trunk extension and flexibility was defined operationally as the distance in inches off the floor the trunk reaches during a trunk lift.

Upper body strength. Upper body strength is essential to daily functioning and becomes increasingly important as an individual ages (CIAR, 1999). Upper body strength was defined operationally as the total number of modified pull-ups completed (no time limit).

Flexibility. Flexibility ensures adequate range of motion at the specific joint and is important to maintaining functionality (CIAR, 1999). Flexibility was defined operationally as the distance in inches reached on the Back Saver Sit-and-Reach test.

### Optimal Reliability

Optimal reliability refers to the theoretical upper limit of test battery reliability (Wood & Safrit, 1984). Optimal reliability will be operationally defined as the first canonical correlation coefficient ( $R_{c1}$ ) and the total redundancy index ( $R_{d \text{ total}}$ ) calculated from the fall test scores intercorrelation matrix and the spring test scores intercorrelation matrix.

## CHAPTER II

### Review of Literature

In the mid-Twentieth Century, Kraus and Hirschland (1954) documented that American children had lower levels of muscular strength and flexibility than their European counterparts. As a result of these findings, President Eisenhower initiated a task force to investigate fitness shortcomings and promote physical fitness among American school children. The result of the task force's deliberations was the establishment of the President's Council on Youth Fitness and the American Alliance for Health, Physical Education, and Recreation (AAHPER) Youth Fitness Test (AAHPER, 1958; 1965). The purpose of this literature review is to address how physical fitness test scores among children have been measured and evaluated, with specific reference to test batteries associated with AAHPERD, since Eisenhower's initiative to improve fitness among American children. The sections of this review are (a) history of fitness testing among school children, (b) objectives of fitness testing in schools, (c) the tracking of health-related fitness scores, and (d) reliability theory.

## **History of Fitness Testing among School Children**

Following the establishment of the President's Council on Youth Fitness, fitness norms for children derived from the Youth Fitness Test (AAHPER, 1958) were established. The AAHPER Youth Fitness Test was used to measure upper and lower body strength, abdominal strength, speed and agility, coordination, and cardiovascular endurance. The first collection of normative data was recorded in the 1957-58 school year and was based on a sample of 8,500 school children in grades 5 through 12. Dr. Paul Hunsicker was the original project director. Normative test scores were categorized by both age and classification (based on age, height, and weight) and the AAHPER norms indicated that American school children had lower levels of fitness compared to children in other countries (AAHPER, 1958; 1965). The original fitness test battery was composed of the following seven components: a) pull-up test (modified pull-up test for girls); b) sit-up test; c) shuttle run (30 ft, four times); d) standing broad jump; e) 50-yd dash; f) softball throw for distance; and g) 600-yd run/walk. A swim test was included but norms were not published.

Hunsicker repeated the survey project in the 1964-65 school year, again at the request of AAHPER. Data were collected on 9,200 school children, grades 5 through 12.

The testing components were exactly the same, with one exception. The 1965 testing procedures allowed girls to substitute a flexed-arm hang (for time) for pull-ups. Again, normative data were published by age and classification. Test component mean scores for boys and girls, across ages, were higher, with the only exception being the softball throw for girls (AAHPER, 1965). The increase in fitness norm scores from the AAHPER surveys may be evidence that fitness scores had improved among American school children. However, scores may have improved due to the familiarity with test items among children and practitioners rather than actual gains in fitness (Blair, 1992).

If the fitness trait did improve from 1958 to 1965, the reward system used by AAHPER could have partially contributed to these fitness gains. Merit, Achievement, and Progress Awards were granted to students who ranked above the 80<sup>th</sup>, 50<sup>th</sup>, and below 50<sup>th</sup> percentiles respectively, in a specified number of components (AAHPER, 1965). The Presidential Fitness Award, initiated in 1966 by the President's Council on Physical Fitness, was a secondary incentive for children to improve their fitness. Students scoring above the 85<sup>th</sup> percentile on all AAHPER test components could receive this award (Stein, 1988).



The American Alliance for Health, Physical Education, and Recreation again modified the Youth Fitness Test battery in 1976 (AAHPER, 1976). The softball throw was eliminated and two optional distance runs were included in the battery. Further, the sit-up protocol changed from a straight leg to flexed-knee position. The 1976 fitness test battery contained the following six tests: (a) pull-up (boys) and flexed arm hang (girls); (b) timed (one minute) flexed knee sit-up; (c) shuttle run; (d) 50-yd dash; (e) standing long jump; and (f) 600 yd-run, one-mile (or 9 minutes) run, or 1.5 miles (or 12 minutes) run. New normative data by age, on 8,500 children in grades 5 through 12, accompanied the test battery. Normative performance percentiles improved among girls but remained relatively unchanged among boys.

In 1980, the American Alliance for Health, Physical Education, Recreation, and Dance (AAHPERD, formerly AAHPER) changed the Youth Fitness Test to measure three components of fitness - cardiorespiratory function, body composition, and abdominal and low back/ hamstring musculoskeletal function (AAHPERD, 1980; Blair, Falls, & Pate, 1983). The emphasis of the 1980 battery was on health-related fitness, as opposed to sport-related performance measured by previous batteries (Falls, Morrow, & Kohl, 1994; Ross &

Gilbert, 1985a; Whitehead, Pemberton, & Corbin, 1989). Physical fitness, at that time, was defined as one's functional capacity to do work that could be improved through activity and reflected by change in test scores (AAHPERD, 1980). The new Health-Related Physical Fitness Test battery was composed of the following test items: (a) one-mile run (or 9 minute run) or 1.5 mile run (or 12 minute run); (b) body fat measure (sum of triceps and subscapular skinfolds, or triceps as a single measure); (c) modified (arms across chest), timed sit-ups; and (d) sit-and-reach test.

Normative standards were established for the new battery from a convenience sample of 12,362 children, aged 6 through 17 years (with the exception of skinfold norms - adopted from the National Health Examination Survey) (AAHPERD, 1984). Normative data were not compared to previous battery norms due to the differences among test items used. The Health-Related Physical Fitness Test was the first battery to promote criterion-reference standards for test performance (AAHPERD, 1984); however, these standards were not widely used.

The Physical Best Test (AAHPERD, 1988) was the next version of the Health-Related Physical Fitness Test. The only modification was in the collection of body composition

scores, where the triceps and calf skinfold measures replaced the triceps and subscapular measures. The first version of the FITNESSGRAM (CIAR, 1987) coincided with the Physical Best Test. AAHPERD dropped the fitness test from Physical Best, redirected the emphasis strictly toward an educational program, and promoted the FITNESSGRAM (CIAR, 1992; 1999) as the test battery to complement Physical Best. The FITNESSGRAM also includes the three health-related components of physical fitness: aerobic capacity, body composition, and muscular strength, endurance, and flexibility.

Atypical of a standard battery, the FITNESSGRAM has enabled practitioners to select various test items to fulfill the three components. The battery calls for: (a) one test of aerobic capacity (one-mile run/walk, Progressive Aerobic Capacity Endurance Run (PACER), or one-mile walk); (b) one test of body composition (percent body fat estimated from the sum of triceps and calf skinfolds or Body Mass Index); (c) abdominal strength test (curl-up); (d) trunk strength test (trunk lift); (e) one test of upper body strength (push-up, pull-up, flexed-arm hang, or modified pull-up); and (g) one test of flexibility (back saver sit-and-reach or shoulder stretch). The biggest change in the FITNESSGRAM from previous test batteries is

the promotion of criterion standards associated with minimal levels of good health, instead of normative standards. Prior to the FITNESSGRAM, criterion-referenced standards were not being utilized with normative data to promote qualitative ratings of fitness test scores (Ross, Pate, Delpy, Gold, & Svilar, 1987). The criterion standards associated with the FITNESSGRAM are the first attempt to document necessary levels of health-related fitness that may yield positive health benefits in adulthood.

#### **Objectives of Fitness Testing in Schools**

The current objective of fitness testing is to help children attain fitness levels, through healthy lifestyle habits, that are sufficient for adequate functioning (ACSM, 1988; Franks, Morrow, & Plowman, 1988). The previously documented fitness test batteries have been used to assess fitness levels of American children over time. Physical fitness testing has also been used in the school setting to screen individuals with inadequate levels of fitness and to promote cognitive learning about habits leading to good health (Pate, 1989). Used in this manner, fitness tests can be used as tools to help children attain educational objectives (Whitehead, Pemberton, & Corbin, 1989).

Seefeldt and Vogel (1989) have indicated that faulty assumptions have influenced the actual definition of physical fitness, and, therefore, have influenced tests intended to measure physical fitness. If inappropriate tests have been used to measure fitness then some contributing attributes of health-related fitness are still unknown. Seefeldt and Vogel noted two major problems with norm-referenced "health-related" fitness test batteries. One, improvements on the tests within a battery have not necessarily generalized to an improvement in health. Two, test batteries have not directly assessed health (e.g., blood pressure). As a result of poor test items or standards, low fitness scores may have had a negative rather than a positive influence on activity habits (Fox & Biddle, 1988; Pate, 1994), thereby reducing habits necessary to attain good health.

In response to Seefeldt and Vogel's stance on faulty assumptions, a consensus has existed in the literature suggesting that current elements of fitness tests (e.g., muscular strength) should indeed be used to evaluate health-related fitness (ACSM, 1988; Franks, Morrow, & Plowman, 1988; Whitehead, Pemberton, & Corbin, 1989). Educationally, fitness tests can serve to identify individual needs for proper programming (Pate, 1989;

Whitehead, Pemberton, & Corbin, 1989), and schools may be the only exposure to constructs of health-related fitness for some children (Fox & Biddle, 1988). Fitness testing serves to identify children with low levels of fitness, to promote habits leading to health, and to motivate children to reach or maintain healthy levels (AAHPERD, 1980). Further, testing can be used to enhance cognitive and affective responses to health-related fitness (Pate, 1994).

Criterion-referenced standards associated with fitness tests may be a possible solution to Seefeldt and Vogel's (1989) concern regarding the relationship between fitness scores and health. The American College of Sports Medicine (ACSM) (1988) has promoted the use of criterion-referenced standards over norm-referenced standards. Criterion-referenced standards represent desirable levels of fitness for the performance of daily tasks and prevention of disease (Cureton, 1994; Morrow, Jackson, Disch, & Mood, 1995; Pate, 1994) and are beneficial for prescription purposes (Safrit & Wood, 1995). From an educational standpoint, Whitehead, Pemberton, and Corbin (1989) have noted that teachers may have had difficulty interpreting normative test scores appropriately. The use of criterion-referenced standards may enhance the interpretation of test scores among children, parents, and teachers as well as

enhance "exercise self-efficacy" among children (Franks, Morrow, & Plowman, 1988; Safrit & Wood, 1995). Although criterion-referenced standards may be theoretically useful, scientific evidence is still needed to ensure that standards have evidence of validity (Cureton, 1994; Pate, 1989; 1994). At this time, specific exercise prescriptions during childhood that yield specific health levels in adulthood are still unknown (Blair & Meredith, 1994).

### **The Tracking of Health-related Fitness Scores**

The most recent set of normative data related to health-related physical fitness was reported from the National Children and Youth Fitness Studies (NCYFS) I (Ross & Gilbert, 1985b) and II (Ross & Pate, 1987b). The NCYFS I comprised the following test items, drawn from the 1980 AAHPERD health-related battery, and norms were established from a national probability sample of 8,800 children, ages 10 through 18 years (grades 5 through 12). Compared to data collected two decades prior by the National Center for Health Statistics (Johnson, Hamill, & Lemshow 1972; 1974), NCYFS I data indicated that children had become fatter (Pate, Ross, Dotson, & Gilbert, 1985; Ross & Gilbert, 1985a).

The NCYFS II was designed to collect health-related fitness data on children ages 6 through 9 (Ross & Pate,

1987a). Tests were similar to those used in the NCYFS I study. Normative data were collected on 4,678 children and indicated higher body fat values, compared to data collected by the National Center for Health Statistics. Indicative of both NCYFS surveys, children have gotten fatter over the past 2 to 3 decades.

Corbin and Pangrazi (1992) suggested that, although body fatness has increased among children over the past 2 decades, body fatness may not be increasing at the same rates among all children. Obese children may be becoming much more obese whereas the same pattern may not be evident among less obese children. Pate, Trost, et al. (1999) have indirectly supported this view by documenting that health-related fitness typically tracks from the elementary ages to middle school ages, especially among children in the highest and lowest health-risk categories. In other words, children at risk for low levels of health-related fitness in grade 5 tend to remain at risk through at least grade 7. Practitioners may need to place more emphasis on children with low levels of health-related fitness in order for these children to attain sufficient levels of health-related fitness.

Although body fatness has increased over the past two decades, Corbin and Pangrazi (1992) and Blair (1992) have



suggested that other health-related measures have not decreased over time. Further, Corbin and Pangrazi indicated that children are able to meet criterion standards for many individual test items on the FITNESSGRAM but that the majority of children cannot pass all test items on the battery (i.e., achieve an overall "healthy" level on the health-related fitness construct). Kuntzleman and Reiff (1992), however, provided longitudinal data that offer evidence of lower aerobic capacity among children over the past 10 years. These authors also suggest that increased body fatness levels are present among all children and not strictly children in the highest quartile. At this point, more empirical evidence is needed to determine changes in health-related fitness over the past several decades as well as to determine specific levels of health-related fitness in childhood that will track to health benefits in adulthood.

### **Reliability Theory**

If fitness tests are used in schools, further reliability and validity evidence is needed to support test use (Cureton, 1994; Seefeldt & Vogel, 1989). Reliability refers to the consistency of a measurement instrument or test (Baumgartner & Jackson, 1995). If a test is reliable, each individual tested should consistently receive the same

indication of his/her true ability on the trait. Classical test theory proposes that an individual's true score on a test is never known, yet it is composed of the individual's observed score and error score (Feldt & McKee, 1958; Safrit, 1976). Within the context of fitness testing, the observed score is the actual score on the test item whereas the true score, or the individual's true ability, is a combination of the observed score minus any measurement error.

A reliability coefficient is an estimate of how accurately a test measures a person's true score or ability (Baumgartner & Jackson, 1995). Reliability, theoretically, is the ratio of true-score variance to observed-score variance, with true score variance yielding no error variance (Baumgartner & Jackson, 1995; Feldt & McKee, 1958; Safrit & Wood, 1995). If an estimated reliability coefficient of .80 is calculated for a push-up test, .80 indicates that the test is 80% accurate at estimating true score variance among subjects (Baumgartner & Jackson, 1995). Further, a reliable test detects true differences among individuals (Safrit, 1976; Safrit & Wood, 1995).

The reliability of physical tests can be estimated by analysis of variance (ANOVA). An ANOVA model is effective for reliability estimation because it enables the

researcher to partition variance from various sources into either true or error variance (Bartko, 1966; Safrit, 1976). The combination of true score variance and error variance yields observed score variance. Safrit (1976) noted that although Fisher introduced the use of the ANOVA model to estimate reliability in 1925, the analysis was not utilized within the physical education field until initiated by Brozek & Alexander (1947) and by Feldt and McKee (1958).

Test-retest (test administered on separate days) or internal consistency (test administered multiple times on the same day) designs are typically the models used to estimate reliability of physical tests (Baumgartner & Jackson, 1995). An intraclass correlation coefficient (ICC) is one possible reliability coefficient that can be calculated from ANOVA. An ICC, for an individual test, represents the consistency of the mean test score for each subject and ranges from 0 (no reliability) to 1 (maximum reliability). A reliability coefficient of 1 indicates that a test perfectly estimates true score variance for every individual in a group (Baumgartner & Jackson, 1995).

An ICC can be estimated from either a one-way or two-way ANOVA. It is important to specify which analysis is used, as the one-way and two-way models define error variance differently (Morrow & Jackson, 1993). A one-way

model incorporates subject variability from the group mean and group mean variability from the grand mean as measurement error whereas a two-way model excludes variability between the group mean and grand mean as error (includes only trial interaction as measurement error) (Bartko, 1966; Haggard, 1958). Although both models have been reported in the literature, a one-way model should be used when the order of scores for a subject are not important (e.g., one week test-retest) and a two-way model should be used when the order of scores is important (e.g., systematic order effect) (Haggard, 1958). Safrit (1976) has recommended the two-way model be used for rater objectivity, yet many studies have included this model in a test-retest design (Appendix A).

In practice, a teacher may collect scores from only one day or trial. The Spearman-Brown prophecy formula, used in conjunction with an ICC, can be calculated from a repeated measures analysis to estimate the reliability of test scores for a single test administration (Baumgartner, 1968; Baumgartner & Jackson, 1995; Safrit & Wood, 1995). The Spearman-Brown analysis yields an adjusted ICC ( $ICC_{adj}$ ) for a theoretical one-trial reliability estimate and can be calculated from a one-way or two-way model.

Norm-Referenced Reliability. The AAHPER/AAHPERD fitness batteries (1958, 1976, 1980) and NCYFS I and II (Ross & Gilbert, 1985b; Ross & Pate, 1987b) were administered to collect normative data on school children. From a reliability theory standpoint, an individual's true mile run score does not change across days/trials when the time between tests is short; therefore, a subject should have the same mile-run score (time) from one week to the next. An ICC or an ICC adjusted for one test/trial (using the Spearman-Brown prophecy formula) gives the practitioner an estimate of how consistently the fitness test measures a trait (e.g., aerobic capacity) across subjects or how effectively a test detected consistent differences among subjects. A high reliability coefficient for a norm-referenced test indicates that the test consistently measured subject ability and detected true differences among subjects (Baumgartner & Jackson, 1995; Safrit & Wood, 1995).

Regarding the test items used in the FITNESSGRAM, adjusted norm-referenced ICC reliability coefficients for 1 trial of the one-mile walk test have been acceptable for third and fourth graders, ranging from .80 to .90, but less stable for younger children (Forbus, 1990; Joyner, 1997; Rikli, Petray, & Baumgartner, 1992). The modified pull-up

norm-referenced test-retest reliability estimates have been as low as .52 (Erbaugh, 1990) and intraclass correlation coefficients adjusted for one trial have ranged from .56 to .91 (Cotton, 1990; Kollath, Safrit, Zhu, & Gao, 1991). Little test-retest reliability evidence has been documented for the curl-up test; however, Forbus (1990) published adjusted ICC coefficients from .85 to .92. The back saver sit-and-reach test, a test of hamstring flexibility, has also yielded high test-retest reliability estimates among children 11-14 years, with adjusted ICC coefficients for one day at .97 (Patterson, et al., 1996).

Criterion-referenced Reliability. Criterion-referenced tests have specific classification standards or criteria (e.g., pass/fail, healthy/unhealthy). Criterion-referenced reliability refers to consistency of classification (Safrit & Wood, 1995). In theory, a child's score on a test should not change from one day to another (test-retest), assuming that the true ability of the child does not change. Therefore, the score should be classified (e.g., healthy) the same from one day to another, assuming the subject's true ability does not change. The FITNESSGRAM (CIAR, 1999) battery uses criterion-referenced standards. These standards represent the minimum levels of an attribute associated with functional health and reduced

risk of disease (Cureton & Warren, 1990). A one-mile run time of 8 minutes for a 15-year-old male is a norm-referenced score, but is interpreted by the FITNESSGRAM criterion standards to indicate a specific level of cardiorespiratory function (i.e., healthy). Proportion of agreement (P), Kappa (K), and Modified Kappa (Kq), instead of an ICC, are estimates of criterion-referenced test reliability (Safrit & Wood, 1995).

Regarding FITNESSGRAM standards, Rikli and colleagues (1992) reported test-retest proportion of agreement values for the one-mile run between .77 and .94 for children between 7 and 9 years of age (grades 2 through 4). Criterion-referenced reliability of the trunk lift has also been high with proportion of agreement values from .93-1.0 and modified Kappa values from .86-1.0 (Jackson, et al., 1996; Patterson, Rethwish, & Wiksten, 1997).

Multivariate Reliability. The aforementioned types of reliability estimates are univariate, rather than multivariate, statistics. Univariate statistics are appropriate for estimating the reliability of individual fitness tests. A canonical correlation analysis (CCA) is a multivariate technique that is suitable for analyzing the relationship between two sets of tests (Thompson, 1984), or specifically a test battery. The total redundancy index

yielded from a CCA represents the strength of the relationship between two sets of variables; the canonical correlation itself represents the relationship between linear composites of variable sets (Thorndike, 1978). Wood and Safrit (1984) proposed a CCA be used to estimate test-retest reliability of physical fitness test batteries. The statistics yielded from a CCA include: (a) an optimum reliability coefficient (ORE) that represents the theoretical optimal reliability of the battery (an evaluation of the first canonical correlation coefficient and the total redundancy index); (b) a total redundancy index that represents the shared variance between two administrations of the battery; and (c) structure coefficients that enable analysis of subtest contribution to the battery (Safrit & Wood, 1987).

Thus far in the literature, the reliability of a test battery typically has not been estimated, but rather the reliability of individual test items has been extended to represent the overall accuracy of the battery (Wood & Safrit, 1984; 1987). The majority of fitness test reliability data have been reported for the individual tests; therefore, previous reliability studies have been conducted with univariate analyses on individual tests. Safrit and Wood (1987) conducted the first multivariate



analysis of battery reliability. The study encompassed an entire middle school student body, ages 11-14 (n=545). Subjects were administered the 1980 Health-Related Physical Fitness Test eight days apart with no practice. Safrit and Wood used a CCA to interpret their data. The canonical correlation reliability coefficients of the test battery for the individual age groups ranged from .75-.81 for males and .76-.84 for females. Safrit and Wood concluded that the Health-Related Physical Fitness Test was a reliable fitness test battery. Dinucci, McCune, and Shows (1990) also used a CCA to determine reliability of a modified version of the Health-Related Physical Fitness battery for college physical education majors. These authors suggested that a modified version also was reliable.

Wood and Safrit (1984) developed a theoretical estimate of battery reliability, the optimal reliability coefficient (ORE). The ORE can be viewed as a theoretical upper limit of test battery reliability or as a reference point to compare actual reliability estimates. The ORE is derived from a canonical correlation analysis of an intercorrelation matrix. The intercorrelation matrix is based on all test scores measured after one administration of the battery. Univariate test-retest coefficients are included on the diagonal of the duplicated intercorrelation

matrix to set up a test-retest model. For a more in-depth discussion of the optimal reliability estimate, see Wood and Safrit (1984).

Reliability and Fitness Testing. Fitness test batteries have continuously changed since the first AAHPERD Youth Fitness Test. Modifications of this original battery have occurred more within testing procedures than within fitness components (i.e., muscular strength and flexibility). Again, if fitness tests are used in schools, further reliability evidence is needed to support test use (Cureton, 1994; Seefeldt & Vogel, 1989), especially when testing protocols are constantly changing. Additional norm-referenced and criterion-referenced reliability evidence is needed to ensure current physical fitness tests are being used appropriately among school children. Further, reliability analysis of a fitness test battery should be conducted from a multivariate rather than a univariate perspective. Therefore, the purpose of this study was to estimate the multivariate test battery reliability of the FITNESSGRAM prior to evaluating relationships between fall and spring health-related fitness scores among elementary school-aged children.

## CHAPTER 3

### Method

The purpose of this study was to determine if the FITNESSGRAM test battery reliably measures norm-referenced health-related physical fitness scores among children in grades 3 through 6. Further, descriptive data were collected to document inter-individual and intra-individual changes in health-related fitness scores among school children.

#### **Subjects**

The population investigated was boys and girls in grades 3 through 6 (ages 8 through 13). Participants were recruited from intact physical education classes at two elementary schools. The investigator sent a permission request letter to the principal (Appendix D), and the school superintendent (Appendix D) asking for permission to perform fitness testing among all students who participated in physical education classes. Because a fitness component was part of the physical education curriculum at each school, a passive consent form was distributed to the parents (Appendix D) and Institutional Review Board approval was obtained prior to testing (Appendix E).

## **Instruments**

The FITNESSGRAM test battery (Cooper Institute of Aerobics Research [CIAR], 1999) was used to measure health-related physical fitness. Documented previously (Appendix A), test items included in the FITNESSGRAM have yielded fair to high univariate reliability estimates but multivariate estimates of the battery have not been obtained. Univariate reliability estimates have been higher among criterion-referenced than norm-referenced data.

The FITNESSGRAM allows the practitioner latitude in choosing the test items to meet the mandated battery components of aerobic capacity, body composition, abdominal strength, trunk extensor strength, upper body strength, and flexibility. The following tests of the FITNESSGRAM battery were used to assess the varying components of health-related physical fitness: (a) The one-mile run/walk (aerobic capacity), (b) the body mass index (body composition); (c) the curl-up (abdominal strength), (d) the trunk lift (trunk extensor strength), (e) the modified pull-up (upper body strength), and (f) the back saver sit-and-reach (flexibility).

## Procedures

Subjects from two schools were tested on two occasions, once in the fall and once in the spring. Each FITNESSGRAM test component was administered in a separate station area during the physical education class. The FITNESSGRAM (CIAR, 1999) procedures for the administration and scoring of all test items were strictly followed. The same seven data collectors scored both the fall and spring scores. Two additional data collectors participated on only one occasion. The testing protocols were unique to each school because class times varied between schools.

School One. In both the fall and spring, testing was conducted over two days within the same week. Class periods lasted 30 minutes, allowing all testing to be completed in two days. On day one, students completed the one-mile run/walk outside. A Rolatape® (Model 300) distance wheel was used to measure a circular distance of 1,056 feet. Students wore identification stickers and completed five laps to finish the aerobic test. One data collector called out the identification number as each student passed the start/finish line and an additional data collector recorded the corresponding number of completed laps and finish time.

Each student returned individually to the gym at the completion of the aerobic test and proceeded to the body mass index (BMI) station. A third data collector recorded shoeless height (in inches) from a wall chart and weight (in pounds) from a stadiometer. Once completed, the student proceeded to the trunk lift station. The principal investigator demonstrated the trunk lift test. Students were then asked to perform the trunk lift as the principal investigator recorded the distance from the floor mat to the chin to the nearest half-inch. Students performed two trials of the trunk lift and then had the opportunity to practice the remaining tests under the supervision of the physical education teacher. This practice session was not standardized in any way.

On Day Two, three testing stations were set up inside the gymnasium. Students were divided into groups of approximately six and each group started at either the back saver sit-and-reach (BSSR), curl-up (CU), or modified pull-up (MPU) station. Students then rotated to the remaining stations at eight-minute intervals.

At the CU station, the group of students formed partner duos and had the opportunity to practice the curl-up prior to testing. During practice, the principal investigator called an up/down cadence at the recommended

rate of one curl every three seconds (20 curl-ups per minute) and helped students correct improper form during the practice. Students then switched places with the partner and partners practiced the CU test. Practice concluded when students had demonstrated the correct form (approximately five curl-ups).

At the conclusion of the practice, the students who practiced first completed the CU test. Students started in a supine position with knees bent, fingertips touching the proximal side of the CU strip. Partners kneeled on either side of the student and held the CU strip in place. When the investigator called the "up" cadence, students attempted to curl-up until their fingertips touched the distal edge of the strip (three inches for students younger than ten years and 4 1/2 inches for students ten years and older). The principal investigator recorded an individual's curl-up score on the second form deviation but continued the cadence until all performing students had finished. Students then switched with their partner and the partner completed the same protocol. At any given time, three to four students were performing the curl-up test.

At the BSSR station, four students watched as two students performed the test on sit-and-reach boxes (as

specified in Appendix A of the FITNESSGRAM Test Administration Manual). Scores were recorded for both the right and left leg. Each student had four trials with each leg extended. The students reached up the BSSR box, with one hand on top of the other, palms down. On the fourth trial, students were asked to hold the reach and a data collector recorded the score (in inches) to the nearest half-inch. Student partners assisted in keeping the extended leg straight by holding the knee down to the floor. Once all students had completed the four trials, students again completed four trials with each leg. The data collector changed the BSSR score only if the score improved.

At the MPU station, a separate data collector demonstrated the test. One student completed the MPU test on the modified pull-up machine (as specified in Appendix A of the FITNESSGRAM Test Administration Manual) as the remaining students waited their turns. In order to adequately place the pull bar and chin strip, students were in a supine position under the pull bar and were asked to reach their arms straight in the air. The data collector then placed the pull bar seven to eight inches above the fingertips and the chin bar was placed four rungs below the pull bar. Students were instructed to complete as many



modified pull-ups as possible. The data collector verbally expressed that the chin should cross over chin strip and that a student's bottom should not touch the ground between pull-ups. Each student completed the test twice and the higher of the two scores was recorded.

School Two. In the fall, testing was conducted for two days each week for two weeks. Class periods lasted 25 minutes and students had not annually been tested on the FITNESSGRAM; hence, more standardized practice was provided than the previous school. Although the order of testing was different between School One and School Two, the testing protocols for the individual tests were identical and the same data collectors collected scores.

In Day One of Week One, students were divided into groups of six to eight students and rotated every 8 minutes among three stations: BMI and trunk lift, BSSR, and CU. At the BMI and trunk lift station, one data collector recorded height, weight, and birth date in a private corridor. A second data collector measured the highest score of two trunk lifts. Students were able to practice the BSSR and CU tests twice under the supervision and correction of the physical education teacher and the principal investigator respectively. No scores were recorded for the BSSR or CU

tests, as these stations served as standardized practice sessions.

On Day Two of Week One, students completed the one-mile run/walk. Students were given identification stickers and completed six laps around a 880 foot outside course. One data collector called out the identification number as each student passed the start/finish line and an additional data collector recorded the corresponding number of completed laps and finish time.

On Day One of Week Two, students were again divided into three groups and rotated at eight minute intervals on three stations: CU, BSSR, and MPU. Official scores were recorded at the CU and BSSR stations by the principal investigator and a data collector respectively, and students practiced the MPU test twice.

On Day Two of Week Two, students were divided into groups and were rotated among the MPU station and two activity stations. Two data collectors recorded the modified pull-up scores (one collector per machine). Each child had two opportunities to perform the test. During Day Two of Week Two, individuals that were absent from previous testing sessions were administered make-ups by the principal investigator. Individuals were also allowed to make up the one-mile run/walk test approximately two weeks

after the original testing. In the spring, students completed the one mile-run on Day One of Week One and all other test items on Day Two.

## Analyses

Optimal Reliability. A canonical correlation analysis was conducted on the entire sample as well as each grade and gender to determine the optimal reliability estimate of the test battery. The optimal reliability estimate is a collective evaluation of the first canonical correlation coefficient ( $R_{c1}$ ) and the total redundancy index ( $R_{d \text{ Total}}$ ) and can be used to determine if the FITNESSGRAM can reliably measure the norm-referenced health-related fitness scores of children in both the fall and spring. Although much reliability data has been reported by age, Safrit and Wood (1987) have indicated that analysis by grade is appropriate since tests are administered by grade.

Inter-individual Changes in Scores. Individual canonical correlation analyses were conducted on the entire sample and for each grade/gender group to determine the stability of health-related fitness scores across an academic school year.

Intra-individual Changes in Scores. Percent agreement statistics were calculated between fall and spring classifications (i.e., healthy/ unhealthy). These statistics were calculated on each test for each grade/gender group.

## Chapter IV

### Results

Data were collected on 390 children, 352 children had complete sets data for the fall and spring. Descriptive statistics for the children's test scores are presented in Table 1. Test scores across age and gender are included in Appendix B.

Table 1

Means and Standard Deviations for the Entire Sample

<u>Test</u>	<u>Fall</u>			<u>Spring</u>		
	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>
TL	10.1	2.1	386	11.0	1.3	378
CU	16.0	11.5	377	19.2	13.6	376
MPU	10.2	7.4	376	9.5	7.0	374
BSSR-L	10.3	1.9	386	10.4	2.0	375
BSSR-R	10.5	1.8	378	10.5	1.8	376
BMI	20.2	4.9	365	20.5	5.1	379
Mile	907	208	365	896	216	376

note. Test items are trunk lift, curl-up, modified pull-up, back saver sit-and-reach left and right, body mass index, and one-mile run/walk respectively. Trunk lift and back saver sit-and-reach scores are reported in inches and the mile time in seconds.

Overall, children tended to improve from fall to spring on most of the test items (Table 1). As documented in Appendix B, boys and girls both improved on the trunk lift and curl-up at each grade level. Boys also improved from fall to spring on the modified pull-up (grades 3 and 4 only) and the one-mile run/walk (grades 3, 4, and 6). Lower levels of performance in the spring were evident for specific grade levels on the modified pull-up (grades 5 and 6, both boys and girls) and the one-mile run (grade 5 for boys and grades 5 and 6 for girls). Flexibility was relatively consistent from fall to spring with minor decreases in the left back saver sit-and-reach for both boys and girls.

### **Optimal Reliability**

The optimal reliability coefficients are presented in Table 2. The first canonical correlation coefficient and the total redundancy index are used in combination to estimate the maximum test battery reliability at a given time period. Again, the optimal reliability is a theoretical estimate only and not a test-retest reliability coefficient.

Table 2

Optimal Reliability Coefficients

	<u>Fall</u>			<u>Spring</u>		
	<u>n</u>	<u>R<sub>cl</sub></u>	<u>R<sub>d Total</sub></u>	<u>n</u>	<u>R<sub>cl</sub></u>	<u>R<sub>d Total</sub></u>
3	67	.982	.79	65	.982	.79
4	91	.982	.82	93	.982	.81
5	99	.982	.80	100	.982	.80
6	95	.982	.81	97	.981	.81
Total	352	.981	.81	355	.981	.81

Based upon two administrations of the test battery, the similar optimal reliability coefficients in the fall and spring are evidence that the FITNESSGRAM test battery reliability is stable at each grade level. The high values associated with the optimal reliability estimates (Table 2) indicate that the battery is also theoretically highly reliable. Although the current estimates are somewhat lower than estimates found in the literature for a similar sample, the current estimates are high enough to reliably measure health-related physical fitness among children in grades 3 through 6. However, the ORE will likely be higher than a test-retest reliability estimate due to additional measurement error associated with the retest session.

### Inter-individual Changes

A canonical correlation analysis (CCA) was used to determine the stability of health-related fitness scores over the course of an academic year. The CCA yields the total redundancy estimate ( $R_{d \text{ total}}$ ) that is an estimate of shared variance between the fall and spring scores, or the variability in one set of scores that can be explained by the variability in the other set of scores.

Table 3

Shared Variance between Fall and Spring Test Battery

Composite Scores for the Entire Sample

	$n$	$R_{c1}$	$R_{d \text{ total}}$
3	67	.948	.48
4	91	.960	.48
5	99	.945	.64
6	95	.987	.57
Total	352	.956	.55



Table 4Shared Variance between Fall and Spring Test BatteryComposite Scores for Boys

	<u>n</u>	<u>R<sub>cl</sub></u>	<u>R<sub>d Total</sub></u>
3	41	.971	.55
4	42	.971	.54
5	50	.978	.67
6	54	.992	.58
Total	187	.975	.58

Table 5Shared Variance between Fall and Spring Test BatteryComposite Scores for Girls

	<u>n</u>	<u>R<sub>cl</sub></u>	<u>R<sub>d Total</sub></u>
3	26	.950	.46
4	49	.966	.41
5	49	.922	.63
6	41	.983	.50
Total	165	.933	.52

Expectedly, the moderate amount of shared variance between fall and spring scores indicates that the health-related fitness of children changes inconsistently during the school year (Tables 3-5). Varying maturity rates and

motivational factors, and not necessarily improvements in health-related fitness, may account for many of the changes in fall to spring scores. If health-related fitness, as a construct, tracked similarly during the year, one would expect to see much higher total redundancy values.

Individuals may improve performance from fall to spring on specific test items; however, the multivariate health-related fitness construct may not improve in the same manner. The purpose of documenting the shared variance statistics from this data set was to establish a baseline of health-related fitness changes (from a multivariate as opposed to univariate perspective) as a marker for future studies.

#### **Intra-individual Changes**

Frequencies (Appendix C) were calculated for both fall and spring scores to determine how many children passed the FITNESSGRAM standards for minimum healthy performance on specific test items. Percent agreement (Tables 6-8) was used to determine the consistency of criterion classification (i.e., healthy, unhealthy) from fall to spring on individual test items. In other words, it was important to determine if children received the same classification during various parts of the year.

Table 6Percent Agreement of Combined Sample on Criterion Standards

	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>T</u>
	<u>PA</u>	<u>PA</u>	<u>PA</u>	<u>PA</u>	<u>PA</u>
MR		87	85	86	86
CU	73	71	71	75	72
PU	78	86	77	78	80
BM	79	85	91	91	87
SR	91	80	89	90	88
TL	90	86	87	90	88

note. MR represents the one-mile run/walk; CU represents the curl-up; PU represents the modified pull-up; BM represents body mass index; SR represents the average of the 2 sit-and-reach scores; TL represents trunk lift.

Table 7Percent Agreement of Boys on Criterion Standards

	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>T</u>
	<u>PA</u>	<u>PA</u>	<u>PA</u>	<u>PA</u>	<u>PA</u>
MR		86	92	84	87
CU	78	77	70	75	75
PU	88	84	76	76	81
BM	86	89	90	87	88
SR	90	86	88	93	90
TL	90	76	87	91	86

Table 8Percent Agreement of Girls on Criterion Standards

	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>T</u>
	<u>PA</u>	<u>PA</u>	<u>PA</u>	<u>PA</u>	<u>PA</u>
MR		90	78	88	83
CU	65	65	71	76	70
PU	63	84	76	80	78
BM	68	81	92	94	85
SR	92	82	90	87	87
TL	100	94	86	89	90

Percents of agreement (from fall to spring) were high for all age/gender groups on all test items. High percent

agreement from fall to spring indicates that children remained relatively stable in their criterion classifications (Tables 6-8). In other words, children who were categorized as healthy in the fall were typically classified as healthy in the spring. On the other hand, children who were classified as unhealthy in the fall typically remained in the unhealthy category.

## CHAPTER V

### Discussion

#### **Optimal Reliability**

The optimal reliability estimate (ORE) is a theoretical evaluation of the highest test-retest reliability of a test battery. In the current study, an ORE was estimated to determine if the FITNESSGRAM battery could reliably measure health-related fitness among elementary-aged school children in both the fall and spring. It is important that test battery reliability remain stable across the school year, as fitness testing may be conducted in both the fall and spring of the same school year for diagnostic or tracking purposes. Again, the ORE is a combination of both the first canonical correlation coefficient ( $R_{c1}$ ) and the total redundancy index ( $R_{d \text{ Total}}$ ). In the current study, ORE estimates (Table 2) were extremely stable in the fall and spring, with total redundancy indexes ranging from .79 to .82 among all grades.

Compared to other ORE estimates reported in the literature, the current sample estimates are slightly lower. Safrit and Wood (1987) reported a  $R_{c1}$  of .98 and a  $R_{d \text{ Total}}$  of .89 for the Health-Related Physical Fitness Test (AAHPERD, 1980) on a sample of children in grades 6 to 8.

Dinucci, McCune, and Shows (1990) reported a  $R_{c1}$  of .99 and a  $R_{d \text{ Total}}$  of .93 for a modified version of the aforementioned battery on a sample of college students. Although the current sample estimates are somewhat lower than other studies, current estimates are certainly acceptable, especially considering multivariate reliability estimates for the FITNESSGRAM are not present in the literature. The lower estimates in the current study could also be due to the younger-aged sample.

### **Inter-individual Changes**

Shared variance between fall and spring health-related fitness scores appear to be moderate. In other words, spring health-related fitness composite scores can only moderately be explained or predicted by fall composite scores. Shared variance statistics ( $R_{d \text{ Total}}$ ) ranged from .48 to .67 (Tables 3-5) across grades. Shared variance statistics were higher among boys than girls, indicating the girls have greater individual or within group variations in either motivation, maturation, or actual fitness changes from fall to spring. Without a control group for comparison purposes, it is difficult to distinguish the effect of maturation on fitness scores, as maturation is associated with greater muscle strength and movement economy. Hence, physical fitness tests may need

to return to norms adjusted for maturation, as in the original Youth Fitness Test (AAHPER, 1958).

Shared variance statistics were highest for children in grade 5. From an anecdotal standpoint, children in grades 3-5 typically appeared to give a maximum effort on all test items. Lower shared variance statistics among children in grades 3 and 4 could be due to the drastic variations in maturity development over the course of an academic year. As documented by Malina (1994), fitness scores may improve due to biological growth (not necessarily changes in fitness). By grade 5, the maturational differences appear to narrow, possibly explaining the higher shared variance between fall and spring scores. Changes in fitness scores among children in grade 5 (Tables 3-5) are less variable than in other grades and may be due more to changes in actual health-related fitness than changes in fitness due to maturation.

Shared variance statistics among children in grade 6, however, are lower than grade 5. Again, from an anecdotal standpoint, children in grade 6 did not always give a maximum effort on all test items, especially the one-mile run/walk test. Although children may have fewer maturational effects on their health-related fitness, motivational factors appear to reduce the explanatory power



of fall health-related fitness scores. In summation, physical education programs may have a larger impact on health-related changes during grades 5 and 6 or maturation differences may be smaller in these grades than grades 3 and 4. Further, fitness scores among children in grade 6 may not be reflective of an individual's true ability due to motivational factors, making it difficult to determine if actual improvements in health-related fitness occurred during the course of a year.

The shared variance statistics from fall to spring health-related fitness scores are novel to the literature. As mentioned previously, the evaluation of fitness tests have predominantly been from a univariate, rather than multivariate, perspective. With the emphasis of physical fitness now geared toward disease prevention and functional capacity maintenance (a multivariate construct), multivariate statistics will likely be utilized more often in the overall evaluation of health-related fitness rather than univariate statistics on a specific trait (e.g., strictly hamstring flexibility). Therefore, it is important to document baseline multivariate statistical data that represent the stability of health-related fitness scores over the course of an academic school year. Using multivariate techniques, researchers will be able to

more fully investigate health-related fitness and how the construct tracks from childhood through adulthood. Future studies incorporating the assessment of health-related fitness as a multivariate construct will now have a reference index in which to gauge fitness changes in elementary-aged school children.

### **Intra-individual Changes**

Children in grades 3 through 6 tended to remain in the same fitness classification (e.g., healthy) in both the fall and spring, as high percent agreement statistics (Tables 6-8) indicate. Similar classification agreement, or tracking, is similar to the longitudinal fitness change results reported by Pate, Trost, et al. (1999). Among all the test items, percent agreement statistics ranged from .75 to .91 among boys. Classification on all items was highly stable across the school year, with the only exception being the curl-up test. Because only modest gains on the mean curl-up score were exhibited in each grade, inconsistent classification could be due to test familiarity (better performance in the spring) or to low standards (modest improvement could result in reclassification). Percent agreement statistics were less stable among the girls, especially on the strength items.

Performance decline on the modified-pull up test could explain the varied classification.

High percent agreement may be viewed as both a positive and a negative. From one vantage point, it is important to document that children participating in physical education class twice a week are capable of maintaining a healthy level of performance on specific traits (i.e., upper body strength). It is somewhat disconcerting, however, that children classified as unhealthy on certain traits typically do not improve during the academic year.

Although percent agreement statistics were high, percentage of children classified as healthy on test items (Appendix C) was not as high. Exceptionally low passing rates were exhibited on the one-mile run/walk test, with passing rates ranging from 10-33% among grade and gender groups in either the fall or spring. Passing rates for the strength test items (curl-up, modified pull-up) were better, ranging from 50-88%. The sit-and-reach and trunk lift passing rates were much higher, ranging from 78-100% (Appendix B). Of concern, however, only 50-60% of children were classified as having healthy levels of body composition. As indicative by research on body fat increases over the past 30 years (Lohman, 1981; Ross &

Gilbert, 1985b; Ross & Pate, 1987a), body fatness is a threat to good health among a disproportionate number of children.

The discrepancy among passing percentages on the different test standards could be due to lower ability on certain traits among this sample or could be an indication that not all test items are contributing equally to the construct of health-related fitness. The low passing percentages on the one-mile run/walk test could indicate that the standards are too difficult to pass or that the sample of children simply may have had poor aerobic capacity. The high passing percentages on the trunk lift and back saver sit-and-reach, however, could be indicative that the standards are too easy or that children have sufficient flexibility for long-term health benefits. Interestingly, a consistent decrease in criterion passing rates occurred between grade 4 and grade 5 on all items except the trunk lift. The decrease may be evidence that the criterion standards for children in grade 5 are too stringent compared to the other grade (age) standards.

#### **Comparison to National Probability Samples**

Compared to the National Child and Youth Fitness Studies (NCYFS) I & II (national probability samples), children in the present sample had considerably lower mean

scores on individual test items. The average one-mile run times for the current sample were between 3 and 5 minutes slower than national averages for each age group and gender. Average upper body strength was also lower in the present sample. The lower scores may indicate that children have lower levels of test-specific fitness than 10-15 years ago (approximate time the NCYFS were conducted) or that local children may not be exposed to sufficient frequency in physical education (only two days per week) to attain average levels of health-related fitness.

It is difficult to compare scores on each test item between the current sample and national probability studies because the test items and protocols have changed during the last decade. For example, the curl-up test has replaced the sit-up test as a measure of abdominal strength. It is possible, however, to determine the percentage of children passing FITNESSGRAM standards on specific traits (e.g., abdominal strength) for both the current sample and national samples.

Looney and Plowman (1990) determined the percentage of children in the NCYFS I and II that passed FITNESSGRAM (1992) criterion standards. Although passing rates on criterion standards were fairly high among the current sample (Appendix B), percentage of passing rates among many

of the test items were substantially lower in the current sample compared to the NCYFS data. Alarming differences were present in both aerobic capacity (60-80% classified as healthy in NCYFS, across ages and genders, compared to 10-33% in the current sample) and body composition (80% across NCYFS age levels compared to 46-70%). Passing rates on the abdominal strength tests, however, were fairly similar. The discrepancy on "healthy" criterion passing rates between the current sample and national samples, despite the lowering of FITNESSGRAM criterion standards, supports the differences detected in the norm-referenced comparisons.

#### **Suggestions for Future Research**

From a measurement perspective, it is important to document the optimal reliability and test-retest reliability of the FITNESSGRAM among various grades. Further, The FITNESSGRAM enables the practitioner to choose the individual test items to measure the construct of health-related fitness. Researchers must now determine what combinations of test items increase the reliability of the instrument and what combinations of items detract from its accuracy. From a health promotion and disease prevention perspective, future research should track criterion scores among individuals from childhood to

adulthood, thereby validating or adjusting FITNESSGRAM criterion standards.

### **Summary**

The FITNESSGRAM test battery appears to be a reliable instrument to measure health-related fitness among children in grades 3 through 6. Children's health-related fitness, as a multivariate construct, is only moderately stable during the academic school year, with maturational changes impacting the stability of scores among children in grades 3 and 4 and motivational factors hampering stability among children in grade 6. Children in grade 5, however, seem to have the most stable health-related fitness scores over time. Finally, although children are consistently classified into the same healthy/unhealthy categories from fall to spring, evidence exists that improvement in health-related fitness may not result from physical education classes only twice per week, especially in terms of body fatness.

## Appendix A

Test-retest Reliability Estimates of Physical Tests  
Included in the FITNESSGRAM



Test-Retest Reliability of the One-Mile Run-Walk

<u>Source</u>	<u>Sex</u>	<u>Grade</u>	<u>r</u>	<u>ICC</u>	<u>ICC<sub>adj</sub></u>
Buono, Roby, Micale, Sallis, & Shepard (1991)	M/F	5		.91	
	M/F	8		.93	
Forbus (1990)	M	11-15 yrs			.82 <sup>a</sup>
	F	11-15 yrs			.80 <sup>a</sup>
Krahenbuhl, et al. (1978)	F	1	.82 <sup>b</sup>		
	M	3	.92		
Rikli, Petray, & Baumgartner (1992)	M	3			.84, .87
	F	3			.90, .87
	M	4			.87, .83
	F	4			.85, .83

Note. Fall ICC values are reported first by Rikli et al., followed by Spring values.

$$^aR = \frac{MS_B - MS_W}{MS_B + MS_W}$$

<sup>b</sup>1600 m run

Criterion-Reference Reliability Coefficients for the One-  
Mile Run/Walk

<u>Source</u>	<u>n</u>	<u>Sex</u>	<u>Grade</u>	<u>PA</u> <u>(fall)</u>	<u>PA</u> <u>(spring)</u>
Rikli, Petray, &	20,13	M	K	.75	.70
Baumgartner (1992)	16,12	F	K	.69	.51
	15,11	M	1	.76	.66
	17,11	F	1	.76	.45
	45,39	M	2	.85	.77
	52,47	F	2	.81	.85
	53,49	M	3	.91	.85
	63,52	F	3	.90	.84
	44,40	M	4	.86	.83
	37,30	F	4	.83	.94

Test-Retest Reliability for Modified Pull-up Tests on  
Children

Source	Sex	Grade	ICC	ICC <sub>adj</sub>
Cotton (1990)	M, F	3	.75, .88	.59, .78
	M, F	4	.90, .92	.82, .86
	M, F	5	.79, .83	.65, .71
	M, F	6	.90, .95	.82, .90
Engleman &	M, F	3	.81, .90	.68, .83
Morrow (1989)	M, F	4	.91, .87	.83, .77
	M, F	5	.87, .90	.77, .82
	M, F	3-5	.87, .89	.77, .81
Jackson, et al. (1982)	M	9-11 yrs		.94
	F			
Kollath, Safrit	M	9		.910
Zhu, & Gao (1991)	F			.721
Pate, Burgess,	M	9-10 yrs		.83
Woods, Ross, &	F			.81
Baumgartner (1993)	M/F			.83

note. The first ICC values listed are for male samples followed by values for female samples.

Test-Retest Reliability of the Curl-up Test

<u>Source</u>	<u>Sex</u>	<u>Subjects</u>	<u>ICC</u>	<u>ICC<sub>adj</sub></u>
Forbus (1990)	M	11-15 yrs		.85 <sup>a</sup>
	F			.92 <sup>a</sup>
Robertson & Magnusdottir (1987)	M	college	.93	
	F		.97	
Robertson & Magnusdottir (1987) - unpublished data	M		.93	
	F		.94	

---


$$^aR = MS_B + MS_W / (MS_B + MS_W)$$

Test-Retest Reliability of the Trunk Lift on Children.

<u>Source</u>	<u>Sex</u>	<u>Subjects</u>	<u>ICC</u>	<u>ICC<sub>adj</sub></u>
Jackson, Morrow, Jensen, Jones, & Schultes (1996)	M	college	.96	.86
	F		.96	.86
Patterson, Rethwisch, & Wiksten (1997)	M	high school	.95, .98 <sup>a</sup>	
	F		.96, .97	
	M		.95 <sup>b</sup>	
	F		.93	

<sup>a</sup>within day reliability

<sup>b</sup>across days reliability

Criterion-Referenced Reliability of the Trunk Lift.

<u>Source</u>	<u>Sex</u>	<u>Subjects</u>	<u>Duration</u>	<u>PA</u>	<u>Kq</u>
Jackson, Morrow, Jensen, Jones, & Schultes (1996)	M	college	1 week	.98	.96
	F			.99	.98
	M/F			.98	.96
Patterson, Rethwisch, & Wiksten (1997)	M	high school	2 days	.93	.86
	F			1.0	1.0

Note. Kq = modified Kappa.

Test-Retest Reliability of the Sit-and-Reach Tests

<u>Source</u>	<u>Sex</u>	<u>Age</u>	<u>Test</u>	<u>ICC</u>	<u>ICC<sub>adj</sub></u>
Forbus (1990) <sup>a</sup>	M	11-15	t		.93
	F	11-15			.93
Jackson & Baker (1986) <sup>b</sup>	F	13-15	t	.99	
Patterson, Wiksten, Ray, Flanders, & Sanphy (1996) <sup>b</sup>	M	11-15	BSSR	.99, .99	.97, .97
	F			.99, .99	.96, .95
Safrit & Wood (1987)	M	11	t	.97	
	F			.93	
	M	12		.97	
	F			.96	
	M	13		.97	
	F			.93	
	M	14		.97	
	F			.89	

Note. Reliability estimates for the back saver sit-and-reach are reported for the left then right side; t represents the traditional test.

<sup>a</sup>test-retest design

<sup>b</sup>internal consistency design

## Appendix B

### Descriptive Statistics on the Individual Test Items



Means and Standard Deviations on the Trunk Lift (in Inches)

		<u>Combined</u>			<u>Boys</u>			<u>Girls</u>		
		<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>
3	f	9.9	2.1	72	9.7	2.1	43	10.1	2.0	29
	s	10.6	1.4	70	10.6	1.4	41	10.6	1.5	29
4	f	9.2	2.3	101	9.0	2.2	47	9.4	2.3	54
	s	10.7	1.6	99	10.5	1.6	47	10.9	1.5	52
5	f	10.5	1.9	105	10.4	1.9	52	10.7	1.9	53
	s	11.2	1.3	102	11.1	1.3	50	11.3	1.3	52
6	f	10.6	1.8	108	10.5	1.9	59	10.6	1.7	49
	s	11.5	0.8	58	11.5	0.8	58	11.5	0.8	49
T	f	10.1	2.1	386	10.0	2.1	201	10.2	2.1	185
	s	11.0	1.3	378	11.0	1.3	196	11.1	1.3	182

note. For tables in Appendix B, the numbers 3-6 represent the grade and T represents all grades collectively. F and s represent fall and spring scores respectively.

Means and Standard Deviations on the Curl-Up

		<u>Combined</u>			<u>Boys</u>			<u>Girls</u>		
		<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>
3	f	8.2	5.9	72	8.4	5.7	43	8.0	6.4	29
	s	10.2	7.4	68	10.6	8.3	41	9.4	5.9	27
4	f	14.5	9.1	99	14.1	8.3	46	14.9	9.7	53
	s	17.4	11.1	97	17.6	11.3	45	17.3	11.1	52
5	f	17.4	12.3	101	16.8	12.2	51	18.0	12.4	50
	s	19.2	13.9	105	17.6	13.5	54	21.0	14.3	51
6	f	21.4	12.7	105	21.8	11.5	56	20.8	14.0	49
	s	26.7	14.3	106	27.9	13.7	60	25.2	15.1	46
T	f	16.0	11.5	377	15.7	11.0	196	16.3	12.1	181
	s	19.2	13.6	376	19.2	13.6	200	19.2	13.6	176

Means and Standard Deviations on the Modified Pull-Up

		<u>Combined</u>			<u>Boys</u>			<u>Girls</u>		
		<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>
3	f	8.0	5.9	72	8.8	6.4	43	6.8	4.8	29
	s	9.6	6.8	68	11.4	7.6	41	6.9	4.0	27
4	f	8.9	5.5	97	9.4	6.0	45	8.5	5.0	52
	s	10.2	7.0	97	11.6	7.1	45	9.0	6.6	52
5	f	9.0	6.6	101	9.4	7.0	51	8.6	6.3	50
	s	7.0	5.8	104	7.4	6.6	53	6.7	4.9	51
6	f	14.0	9.0	106	16.4	10.4	57	11.2	6.0	49
	s	11.3	7.6	105	13.6	8.3	59	8.4	5.5	46
T	f	10.2	7.4	376	11.3	8.4	196	9.0	5.8	180
	s	9.5	7.0	374	11.0	7.8	198	7.9	5.5	176

Means and Standard Deviations on the Left Back Saver Sit-  
and-Reach (in Inches)

		<u>Combined</u>			<u>Boys</u>			<u>Girls</u>		
		<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>
3	f	10.3	1.8	72	10.1	1.9	43	10.6	1.6	29
	s	10.4	2.1	68	9.9	2.3	41	11.1	1.5	27
4	f	10.3	1.7	99	10.0	1.7	46	10.5	1.7	53
	s	10.1	2.4	101	10.0	2.1	45	10.7	1.6	52
5	f	10.1	2.4	101	9.5	2.5	51	10.8	2.2	50
	s	10.3	2.3	105	9.9	2.5	54	10.7	2.1	51
6	f	10.6	1.6	106	10.0	1.7	57	11.2	1.2	49
	s	10.6	1.5	105	10.2	1.6	60	11.1	1.2	45
T	f	10.3	1.9	386	9.9	2.0	197	10.8	1.7	181
	s	10.4	2.0	375	10.0	2.1	200	10.9	1.6	175

Means and Standard Deviations the Right Back Saver Sit-and-Reach (in Inches)

		<u>Combined</u>			<u>Boys</u>			<u>Girls</u>		
		<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>
3	f	10.5	1.9	72	10.3	2.1	43	10.9	1.6	29
	s	10.6	1.8	68	10.2	1.9	41	11.1	1.6	27
4	f	10.6	1.7	99	10.2	1.9	46	10.9	1.4	53
	s	10.5	1.8	97	9.8	2.2	45	11.0	1.2	52
5	f	10.4	2.0	101	9.8	2.3	51	10.9	1.6	50
	s	10.5	2.2	102	10.0	2.3	54	11.0	1.8	51
6	f	10.6	1.7	106	9.9	2.0	57	11.2	1.1	49
	s	10.6	1.5	106	10.3	1.5	60	11.1	1.3	46
T	f	10.5	1.8	378	10.1	2.0	197	11.0	1.4	181
	s	10.5	1.8	376	10.1	2.0	200	11.0	1.5	176

Means and Standard Deviations on the Body Mass Index

		<u>Combined</u>			<u>Boys</u>			<u>Girls</u>		
		<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>
3	f	18.6	3.7	72	18.6	3.8	43	18.4	3.5	29
	s	18.9	4.0	72	18.9	4.0	43	18.9	4.0	29
4	f	18.9	4.0	101	19.2	4.1	47	18.7	4.0	54
	s	19.2	4.7	100	19.6	4.0	47	19.3	4.6	53
5	f	21.8	5.7	105	22.3	5.8	52	21.4	5.7	53
	s	22.1	5.5	103	22.7	5.6	51	21.5	5.5	52
6	f	20.8	4.9	107	21.0	5.0	59	20.6	4.8	48
	s	21.1	5.0	104	21.5	5.3	57	20.7	4.6	47
T	f	20.2	4.9	365	20.4	5.0	201	19.9	4.8	184
	s	20.5	5.1	379	20.8	5.0	198	20.2	4.9	181

Means and Standard Deviations on the One-Mile Run/Walk (in Seconds)

	<u>Combined</u>			<u>Boys</u>			<u>Girls</u>			
	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>n</u>	
3	f	957	193	67	935	197	41	993	184	26
	s	928	171	71	902	189	43	969	133	28
4	f	929	213	95	899	231	44	955	194	51
	s	901	197	96	855	205	46	943	182	50
5	f	938	204	104	942	200	52	933	210	52
	s	954	254	102	957	266	51	951	243	51
6	f	819	194	99	782	194	58	870	184	41
	s	815	198	107	729	174	58	916	176	49
T	f	907	208	365	883	215	195	934	198	170
	s	896	216	376	855	228	198	942	193	178

## Appendix C

### Frequency of Healthy Classification of FITNESSGRAM Standards



Frequency (%) of Healthy Classification on FITNESSGRAM

Standards for the One-mile Run/walk

	<u>Fall</u>			<u>Spring</u>		
	<u>B</u>	<u>G</u>	<u>T</u>	<u>B</u>	<u>G</u>	<u>T</u>
4	24 (5)	10 (1)	19 (6)	30 (11)	19 (6)	25 (17)
5	12 (6)	21 (11)	17 (17)	10 (5)	22 (11)	16 (16)
6	28 (16)	20 (8)	24 (24)	33 (19)	14 (7)	21 (26)
Total	21 (27)	19 (20)	20 (47)	24 (35)	18 (24)	21 (59)

note. Numbers in parentheses represent the actual number of children who passed the standard.

Frequency (%) of Healthy Classification on FITNESSGRAM

Standards for the Curl-Up

	<u>Fall</u>			<u>Spring</u>		
	<u>B</u>	<u>G</u>	<u>T</u>	<u>B</u>	<u>G</u>	<u>T</u>
3	54 (23)	50 (14)	52 (37)	56 (23)	56 (15)	56 (38)
4	70 (32)	62 (33)	66 (65)	71 (32)	69 (36)	70 (68)
5	60 (30)	56 (28)	58 (58)	49 (26)	61 (31)	55 (57)
6	70 (39)	51 (25)	61 (64)	75 (44)	61 (28)	69 (72)
Total	64 (124)	56 (100)	60 (224)	63 (125)	63 (110)	63 (235)

Frequency (%) of Healthy Classification on FITNESSGRAM

Standards for the Modified Pull-Up

	<u>Fall</u>			<u>Spring</u>		
	<u>B</u>	<u>G</u>	<u>T</u>	<u>B</u>	<u>G</u>	<u>T</u>
3	86 (37)	76 (22)	82 (59)	88 (36)	71 (19)	81 (55)
4	82 (37)	85 (44)	84 (81)	87 (39)	83 (43)	85 (82)
5	66 (33)	84 (42)	75 (75)	52 (27)	75 (38)	63 (65)
6	86 (49)	94 (46)	90 (95)	78 (45)	76 (35)	77 (80)
Total	80 (156)	86 (154)	83 (310)	75 (147)	77 (135)	76 (282)

Frequency (%) of Healthy Classification on FITNESSGRAM

Standards for the Trunk Lift

	<u>Fall</u>			<u>Spring</u>		
	<u>B</u>	<u>G</u>	<u>T</u>	<u>B</u>	<u>G</u>	<u>T</u>
3	95 (40)	100 (28)	97 (58)	93 (38)	90 (25)	91 (54)
4	83 (38)	89 (47)	86 (85)	89 (42)	90 (47)	90 (89)
5	84 (43)	83 (44)	84 (87)	94 (45)	98 (50)	96 (95)
6	88 (50)	89 (42)	89 (92)	100 (57)	100 (48)	100 (105)
Total	87 (171)	89 (161)	88 (332)	94 (182)	95 (171)	95 (353)

Frequency (%) of Healthy Classification on FITNESSGRAM

Standards for the Body Mass Index

	<u>Fall</u>			<u>Spring</u>		
	<u>B</u>	<u>G</u>	<u>T</u>	<u>B</u>	<u>G</u>	<u>T</u>
3	70 (30)	57 (16)	65 (46)	70 (30)	48 (14)	61 (44)
4	60 (28)	54 (29)	56 (57)	62 (29)	44 (23)	53 (52)
5	46 (24)	55 (29)	51 (53)	42 (21)	60 (31)	51 (52)
6	61 (36)	67 (32)	64 (68)	48 (27)	70 (33)	58 (60)
Total	59 (118)	58 (106)	58 (224)	55 (107)	56 (101)	55 (208)

Frequency (%) of Healthy Classification on FITNESSGRAM

Standards for the Combined Back Saver Sit-and-Reach

	<u>Fall</u>			<u>Spring</u>		
	<u>B</u>	<u>G</u>	<u>T</u>	<u>B</u>	<u>G</u>	<u>T</u>
3	81 (35)	86 (24)	83 (59)	85 (35)	93 (25)	88 (60)
4	89 (41)	85 (45)	87 (86)	87 (39)	90 (47)	89 (86)
5	77 (39)	84 (42)	80 (81)	85 (46)	78 (40)	82 (86)
6	86 (49)	92 (45)	89 (94)	95 (57)	82 (37)	90 (94)
Total	83 (164)	87 (156)	85 (320)	89 (177)	85 (149)	87 (326)

## Appendix D

### Approval Letters for School Personnel

Principal, Black Fox Elementary

Murfreesboro, TN 37130

I am a doctoral student at Middle Tennessee State University and am conducting dissertation research on measurement issues regarding the Prudential FITNESSGRAM physical fitness test battery. I would like to test students at Black Fox Elementary in order to determine how accurately this test measures health-related physical fitness over time. The physical education teacher, in order to measure fitness of children during the typical school year, currently uses the test battery protocol.

All test results will be confidential. I will only use the test scores to perform the appropriate analyses. At that point, all names will be destroyed. The Prudential FITNESSGRAM has established test standards that are indicative of minimum levels of functioning necessary for good health. The children will benefit directly from testing by receiving feedback on healthy and unhealthy performance levels. The data analyses can be directly beneficial to the teacher, allowing an estimation of the test battery accuracy in determining healthy and unhealthy levels of fitness.



Because testing is already part of each student's physical education experience, the Code of Federal Regulations (Title 34 - Education, Part 97 - Protection of Human Subjects) does not mandate that student or parent permission is obtained. I will, however, notify parents (enclosed) beforehand that I will be collecting test scores and seek student assent via Mr. Vaughn (enclosed). If a parent contacts me (my telephone number is on the notice to parents) or Mr. Vaughn, I will not record his/her child's test scores. I have already received consent from Mr. Vaughn. I am also seeking consent from the Director of Schools. If you will allow scores to be collected, please sign the letter of approval (enclosed). Thank you for your time. Please do not hesitate to contact me with any questions. I can be reached at 898-5545.

Sincerely,

J.P. Barfield

**LETTER OF APPROVAL - PRINCIPAL**

Department of Health, Physical Education, Recreation, and  
Safety

Middle Tennessee State University

---

Principal Investigator

---

Responsible Faculty Member

**Project Title: Stability Reliability of the Prudential  
FITNESSGRAM among Children in Grades 3-6.**

Please indicate below if you understand the scope and purpose of the research project and give your consent for data collection. Please return in the enclosed envelope or fax (898-5020) by September 17, 1999.

**I CERTIFY THAT I HAVE READ AND UNDERSTOOD THE ABOVE  
RESEARCH PROJECT. I WILLINGLY CONSENT TO THE COLLECTION OF  
TEST SCORES AT BLACK FOX SCHOOL.**

---

Signature of Principal

---

Date

Ms. Marilyn Mathis, Director of Schools, Murfreesboro City Schools

I am a doctoral student at Middle Tennessee State University and am conducting dissertation research on measurement issues regarding the Prudential FITNESSGRAM physical fitness test battery. I would like to test students at Bellwood Elementary and Black Fox Elementary in order to determine how accurately this test measures health-related physical fitness over time. The physical education teacher to measure fitness of children during the typical school year currently uses the test battery protocol.

All test results will be confidential. I will only use the test scores to perform the appropriate analyses. At that point, all names will be destroyed. The Prudential FITNESSGRAM has established test standards that are indicative of minimum levels of functioning necessary for good health. The children will benefit directly from testing by receiving feedback on health and unhealthy performance levels. The data analyses can be directly beneficial to the teacher, allowing an estimation of the test battery accuracy for determining healthy and unhealthy levels of fitness.

Because testing is already part of each student's physical education experience, the Code of Federal Regulations (Title 34 - Education, Part 97 - Protection of Human Subjects) does not mandate that student or parent permission is obtained. I will, however, notify parents (enclosed) beforehand that I will be collecting test scores and seek student assent via the physical education teacher (enclosed). If a parent contacts me (my telephone number is on the notice to parents) or the physical education teacher at the respective school, I will not record his/her child's test scores. I have already received consent from the physical education teacher at each school (Ms. Tina Hall - Bellwood; Mr. Vaughn - Black Fox). I am also seeking consent from the principal at each school (Mr. Joe Thompson - Bellwood; Mr. Zane Cantrell - Black Fox). If you will allow scores to be collected, please sign the letter of approval (enclosed). Thank you for your time. Please do not hesitate to contact me with any questions. I can be reached at 898-5545.

Sincerely,

**J.P. Barfield**

**LETTER OF APPROVAL - DIRECTOR OF SCHOOLS**

Department of Health, Physical Education, Recreation, and  
Safety

Middle Tennessee State University

\_\_\_\_\_  
Principal Investigator

\_\_\_\_\_  
Responsible Faculty Member

**Project Title: Stability Reliability of the Prudential  
FITNESSGRAM among Children in Grades 3-6.**

Please indicate below if you understand the scope and purpose of the research project and give your consent for data collection. Please return in the enclosed envelope or fax (898-5020) by September 17, 1999.

**I CERTIFY THAT I HAVE READ AND UNDERSTOOD THE ABOVE  
RESEARCH PROJECT. I WILLINGLY CONSENT TO THE COLLECTION OF  
TEST SCORES AT BELLWOOD AND BLACK FOX SCHOOL.**

\_\_\_\_\_  
Signature of Director of Schools

\_\_\_\_\_  
Date

Oral Script to Students Read by the Physical Education  
Teacher

Next week, we'll start fitness testing. This year I will have a helper - his name is J.P. J.P. would also like to record your test scores to determine how well the tests work. If you do not want J.P. to record your test score, it is OK. I will not think any differently of you if you do not want him to record your scores. You can tell me before or during testing, in private, if you do not want J.P. to record your scores.

## Note to Parents

Dear Parents and Guardians,

Your child will be participating in fitness testing during the week of October 18th. Fitness tests allow your child to understand if he/she is in good shape. I will be assisting the physical education teacher. I will also be recording your child's scores for research purposes. The research pertains to the accuracy of the tests that the physical education teacher uses.

All names will be kept confidential. Once the research has been completed, all names will be destroyed. Your child's scores will not be distributed to anyone other than two assistants, the physical education teacher, and myself. If you do not want me to record your child's score, your child's grade will not be affected by your decision. If you DO NOT wish for your child's score to be recorded, please sign at the bottom and return to the physical education teacher by October 18, 1999. Please understand that your child will be allowed to participate unless you deny permission.

Thank you.

Sincerely,

**J.P. Barfield**

Sign Below if you DO NOT want the researcher to record your  
child's score.

\_\_\_\_\_  
Parent's Signature

\_\_\_\_\_  
Child's Name

\_\_\_\_\_  
Date



Appendix E  
Institutional Review Board Approval

on-campus memo:



To: J. P. Barfield

From: Nancy Bertrand *Nancy Bertrand*  
IRB Representative  
Nancy Bertrand *Nancy Bertrand*  
IRB Representative

Re: "Stability Reliability of the Prudential FitnessGram  
among Children in Grades 3-6"  
(IRB Protocol Number: 00-014)

Date: November 16, 1999

Thank you for supplying the approval form from the Nashville Metropolitan Schools.

The above named human subjects research proposal has been reviewed and approved. This approval is for one year only. Should the project extend beyond one year or should you desire to change the research protocol in any way, you must submit a memo describing the proposed changes or reasons for extensions to your college's IRB representative for review.

Best of luck in the successful completion of your research.

cc: Dr. David Rowe

## REFERENCES

American Association for Health, Physical Education, and Recreation. (1958). Youth fitness test manual for the national physical fitness test program. Washington, DC:

Author.

American Association for Health, Physical Education, and Recreation. (1965). AAHPER youth fitness test manual (Revised Ed.). Washington, DC: Author.

American Association for Health, Physical Education, and Recreation. (1976). Youth fitness testing manual. Washington, DC: Author.

American Alliance for Health, Physical Education, Recreation, and Dance. (1980). Health-related physical fitness test manual. Washington, DC: Author.

American Alliance for Health, Physical Education, Recreation, and Dance. (1984). Health-related physical fitness technical manual. Washington, DC: Author.

American Alliance for Health, Physical Education, Recreation, and Dance. (1988). Physical Best. Washington, DC: Author.

American College of Sports Medicine. (1988). Opinion statement on physical fitness in children and youth.

Medicine and Science in Sports and Exercise, 20,422-423.

American College of Sports Medicine. (1995). ACSM guidelines for exercise testing and prescription (5<sup>th</sup> ed.). Baltimore: Williams & Wilkins.

Bartko, J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

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

Baumgartner, T. & Jackson, A. (1995). Measurement for evaluation in physical education and exercise science. (5<sup>th</sup> Edition). Dubuque, IO: Brown & Benchmark

Buono, M., Roby, J., Micale, F., Sallis, J., & Shepard, E. (1991). Validity and reliability of predicting maximum oxygen uptake via field tests in children and adolescents. Pediatric Exercise Science, 3, 250-255.

Blair, S. (1992). Are American children and youth fit? The need for better data. Research Quarterly for Exercise and Sport, 63, 120-123.

Blair, S., Falls, H., & Pate, R. (1983). A new physical fitness test. The Physician and Sports Medicine, 11(4), 87-95.

Blair, S., & Meredith, M. (1994). The exercise-health relationship: Does it apply to children and youth? In R. Pate & R. Hohn (Eds.), Health and fitness through physical education (pp. 11-19). Champaign, IL: Human Kinetics.

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

Cooper Institute for Aerobics Research. (1987). The Prudential FITNESSGRAM Test Administration Manual. Dallas, TX: Author.

Cooper Institute for Aerobics Research. (1992). The Prudential FITNESSGRAM Test Administration Manual. Dallas, TX: Author.

Cooper Institute for Aerobics Research. (1999). The FITNESSGRAM Test Administration Manual (2<sup>nd</sup> edition). Champaign, IL: Human Kinetics.

Corbin, C., & Pangrazi, R. (1992). Are American children and youth fit? Research Quarterly for Exercise and Sport, 63, 96-106.

Cotton, D. (1990). An analysis of the NCYFS II modified pull-up. Research Quarterly for Exercise and Sport, 61, 272-274.

Cureton, K. (1994). Physical fitness and activity standards. In R. Pate & R. Hohn (Eds.), Health and fitness through physical education (pp. 129-136). Champaign, IL: Human Kinetics.

Cureton, K. & Warren, G. (1990). Criterion-referenced standards for youth health-related fitness tests: A tutorial. Research Quarterly for Exercise and Sport, 61, 7-19.

Dinucci, J., McCune, D., & Shows, D. (1990). Reliability of a modification of the health-related physical fitness test for use with physical education majors. Research Quarterly for Exercise and Science, 61, 20-25.

Engelman, M., & Morrow, J. (1991). Reliability and skinfold correlates for traditional and modified pull-ups in children grades 3-5. Research Quarterly for Exercise and Sport, 62, 88-91.

Erbaugh, S. (1990). Reliability of physical fitness tests administered to young children. Perceptual and Motor Skills, 71, 1123-1128.

Falls, H., Morrow, J., & Kohl, H. (Eds.). (1994). The Prudential FITNESSGRAM Technical Reference Manual. Dallas, TX: Cooper Institute for Aerobics Research.

Feldt, L., & McKee, M. (1958). Estimation of the reliability of skill tests. Research Quarterly, 29, 279-293.

Forbus, W. R. (1990). The suitability and reliability of the Physical Best fitness test with selected special populations (Unpublished doctoral dissertation, University of Georgia, Athens). Microform Publications, University of Oregon.

Fox, K., & Biddle, S. (1988). The use of fitness tests: Educational and psychological considerations. Journal of Physical Education, Recreation, and Dance, 59(2), 47-53.

Franks, D., Morrow, J., & Plowman, S. (1988). Youth fitness testing: Validation, planning, and politics. Quest, 40, 187-199.

Haggard, E. (1958). Intraclass correlation and the analysis of variance. New York: Dryden Press.

Jackson, A., & Baker, A. (1986). The relationship of the sit and reach test to criterion measures of hamstring and back flexibility in young females. Research Quarterly for Exercise and Sport, 57, 183-186.

Jackson, A., Bruya, L., Baun, W., Richardson, P. Weinberg, R., & Caton, I. (1982). Baumgartner's modified

pull-up test for male and female elementary school aged children. Research Quarterly, 53, 163-164.

Jackson, A., Morrow, J., Jensen, R., Jones, N., & Schultes, S. (1996). Reliability of the prudential FITNESSGRAM trunk lift test in young adults. Research Quarterly for Exercise and Sport, 67, 155-177.

Johnson, F., Hamill, D., & Lemshow, S. (1972). Skinfold Thickness of Children 6-11 years. (Series II, No. 120). Washington, DC: U.S. Center for Health Statistics.

Johnson, F., Hamill, D., & Lemshow, S. (1974). Skinfold Thickness of Youth 12-17 years. (Series II, No. 132). Washington, DC: U.S. Center for Health Statistics.

Joyner, B. (1997). Reliability of criterion-referenced standards of the FITNESSGRAM PACER. G.A.H.P.E.R.D. Journal, 31(3), 14-15.

Kollath, J.A., Safrit, M., Zhu, W., & Gao, L-g. (1991). Measurement errors in modified pull-ups testing. Research Quarterly for Exercise and Sport, 62, 432-435.

Krahenbuhl, G., Pangrazi, R., Petersen, G., Burkett, L, & Schneider, M. (1978). Field testing of cardiorespiratory fitness in primary school children. Medicine and Science in Sport and Exercise, 10, 208-213.



Kraus, H., & Hirschland, R. P. (1954). Minimum muscular fitness tests in school children. Research Quarterly for Exercise and Sport, 25, 178-188.

Kuntzleman, C., & Reiff, G. (1992). The decline in American children's fitness scores. Research Quarterly for Exercise and Sport, 63, 107-111.

Lohman, T. (1981). Skinfolds and body density and their relation to body fatness: A review. Human Biology, 53, 181-225.

Looney, M., & Plowman, S. (1990). Passing rates of American children and youth on the FITNESSGRAM criterion-referenced physical fitness standards. Research Quarterly, 61, 215-233.

Malina, R. (1994). Physical activity: Relationship to growth, maturation, and physical fitness. In C. Bouchard, R. Shephard, & T. Stephens (Eds.), Physical activity, fitness, and health (pp. 918-930). Champaign, IL: Human Kinetics.

Morrow, J. (1992). Are American children and youth fit? Review and commentary. Research Quarterly for Exercise and Sport, 63, 95.

Morrow, J., & Jackson, A. (1993). How "significant" is your reliability? Research Quarterly for Exercise and Sport, 64, 352-255.

- Morrow, J., Jackson, A., Disch, J., & Mood, D. (1995). Measurement and Evaluation in Human Performance. Champaign, IL: Human Kinetics.
- Pate, R. (1983). A new definition of youth fitness. The Physician and Sports Medicine 11(4), 77-83.
- Pate, R. (1989). The case for large-scale physical fitness testing in American youth. Pediatric Exercise Science, 1, 290-294.
- Pate, R. (1994). Fitness testing: Current approaches and purposes in physical education. In R. Pate & R. Hohn (Eds.), Health and fitness through physical education (pp. 119-127). Champaign, IL: Human Kinetics.
- Pate, R., Burgess, M., Woods, J., Ross, J., & Baumgartner, T. (1993). Validity of field tests of upper body muscular strength. Research Quarterly for Exercise and Sport, 64, 17-24.
- Pate, R., Ross, J., Dotson, C., & Gilbert, G. (1985). The new norms: A comparison with the 1980 AAHPERD norms. JOPERD, 56(1), 28-30.
- Pate, R., & Shephard, R. (1989). Characteristics of physical fitness in youth. In C. Gisolfi, & D. Lamb (Eds.), Perspectives in Exercise Science and Sports Medicine: Vol 2. Youth, Exercise, & Sport (pp. 1-45). Indianapolis, IN: Benchmark Press.

Pate, R., Trost, S., Dowda, M., Ott, A., Ward, D., Saunders, R., & Felton, G. (1999). Tracking of physical activity, physical inactivity, and health-related physical fitness in rural youth.

Patterson, P., Rethwish, N., & Wiksten, D. (1997). Reliability of the trunk lift in high school boys and girls. Measurement in physical education and exercise science, 1(1), 145-151.

Patterson, P., Wiksten, D., Ray, L, Flanders, C., & Sanphy, D. (1996). The validity and reliability of the back saver sit-and-reach in middle school girls and boys. Research Quarterly for Exercise and Sport, 67, 448-451.

Rikli, R., Petray, C., & Baumgartner, T. (1992). The reliability of distance run tests for children in grades K-4. Research Quarterly for Exercise & Sport, 63, 270-276.

Robertson, L, & Magnusdottir, H. (1987). Evaluation of criteria associated with abdominal fitness testing. Research Quarterly for Exercise and Sport, 58, 355-349.

Ross, J., & Gilbert, G. (1985a). A summary of findings. Journal of Physical Education, Recreation, and Dance, 56(1), 51-53.

Ross, J., & Gilbert, G. (1985b). The national children and youth fitness study [monograph]. Journal of Physical Education, Recreation, and Dance, 56(1), 44-89.

Ross, J., & Pate, R. (1987a). A summary of findings. Journal of Physical Education, Recreation, and Dance, 58(9), 51-56.

Ross, J., & Pate, R. (1987b). The national children and youth fitness study [monograph]. Journal of Physical Education, Recreation, and Dance, 58(9), 50-95.

Ross, J., Pate, R., Delpy, L., Gold, R., & Svilar, M. (1987). New health-related fitness norms. Journal of Physical Education, Recreation, and Dance, 58(9), 66-70.

Safrit, M. (Ed.). (1976). Reliability theory. Washington, DC: AAHPERD.

Safrit, M. (1990). The validity and reliability of fitness tests for children: A review. Pediatric Exercise Science, 2, 9-28.

Safrit, M., & Wood, T. (1987). The test battery reliability of the health related physical fitness test. Research Quarterly for Exercise and Sport, 58, 160-167.

Safrit, M., & Wood, T. (1995). Introduction to measurement in physical education and exercise science. (3<sup>rd</sup> Edition). St. Louis, MO: Mosby.

Seefeldt, V., & Vogel, P. (1989). Physical fitness testing of children: A 30-year history of misguided efforts? Pediatric Exercise Science, 1, 295-302.

Stein, J. (1988). Physical fitness testing and rewards. Journal of Physical Education, Recreation, and Dance, 59(1), 53-57.

Thompson, B. (1984). Canonical Correlation Analysis: Uses and Interpretation. London: Sage Publications.

Thorndike, R. (1978). Correlational procedures for research. New York: Gardner Press.

Updyke, W. (1992). In search of relevant and credible physical fitness standards for children. Research Quarterly for Exercise and Sport, 63, 112-119.

Whitehead, J., Pemberton, C., & Corbin, C. (1989). Perspectives on the physical fitness testing of children: The case for a realistic educational approach. Pediatric Exercise Science, 2, 111-123.

Wood, T., & Safrit, M. (1984). A model for estimating the reliability of psychomotor test batteries. Research Quarterly for Exercise and Science, 55, 53-63.

Wood, T., & Safrit, M. (1987). A comparison of three multivariate models for estimating test battery reliability. Research Quarterly for Exercise and Sport, 58, 150-159.