

## 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.

ProQuest Information and Learning  
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA  
800-521-0600

UMI<sup>®</sup>



**Criterion-Referenced Agreement of the FITNESSGRAM Upper-Body Tests of Muscular  
Strength and Endurance**

**Todd Sherman**

**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 2001**

UMI Number: 3016337

UMI<sup>®</sup>

---

UMI Microform 3016337

Copyright 2001 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

Criterion-Referenced Agreement of the FITNESSGRAM Upper-Body Tests of Muscular  
Strength and Endurance

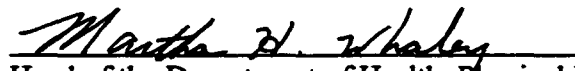
APPROVED:

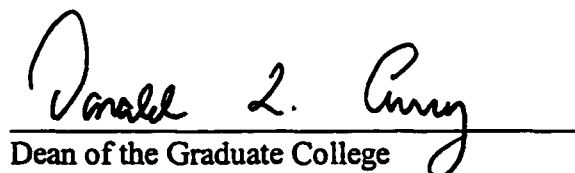
Graduate Committee:

  
Major Professor

  
Committee Member

  
Committee Member

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

  
Dean of the Graduate College

## ABSTRACT

### Criterion-Referenced Agreement of the FITNESSGRAM Upper-Body Tests of Muscular Strength and Endurance

The purpose of the study was to investigate the percent agreement between the FITNESSGRAM push-up test (PSU) and the FITNESSGRAM alternate tests of upper-body strength and endurance. Further, the upper-body strength performances were compared across age groups and genders using survival analysis techniques.

Four hundred and three children, in grades three through six, from a local elementary school were recruited for the study. On the first day of data collection the children's height and weight were measured and the modified pull-up test (MPU) and the flexed-arm hang test (FAH) were administered. On the second day, children were administered the pull-up test (PU) and the push-up test.

The percent agreement indices for eight and eleven-year-old boys were moderate to high (.61 to .86). The PSU-MPU and PSU-FAH percent agreement indices were higher than the PSU-PU percent agreement index for eight, nine, and eleven-year-old boys. The kappa and modified kappa statistics for all three comparisons indicated a slight to substantial agreement (.28-.70 and .22-.72 respectively).

Eight to eleven-year-old girls yielded higher percent agreement indices for the PSU-PU comparisons (.67 to .82) than the PSU-MPU and PSU-FAH (.48 to .75) comparisons. The kappa and modified kappa statistics ranged from .09 to .55 and -.04 to .64 indicating a poor to moderate agreement. Overall, eight to eleven-year-old boys had

higher percent agreement indices than eight to eleven-year-old girls on all three comparisons. The same was indicated for kappa and modified kappa.

As for the survival analyses, the four tests of upper-body strength and endurance did not statistically differentiate ( $p >.05$ ) the strength differences that are typically seen in boys and girls from age group to subsequent age group. On the other hand, strength and endurance levels between boys and girls were statistically different ( $p <.05$ ), with the boys' strength and endurance levels being higher. Those strength differences only held up for the push-up test across all age groups.

In conclusion, based on the large number of poor to moderate agreement indices, using the FITNESSGRAM alternative tests of upper-body strength and endurance will result in different healthy/unhealthy classifications for a high percentage of children, especially girls. Further, a longitudinal study needs to be conducted to compare survival curves across time to assess changes in children's muscular fitness performances.

## ACKNOWLEDGMENTS

To my parents, I will always cherish your wisdom, patience, support, and unconditional love.

I would like to thank my committee for guiding me through the process. Dr. Wagoner, you truly stand for what academia ought to represent.

Finally, I would like to thank all those who helped me collect data.



## TABLE OF CONTENTS

	Page
List of Tables .....	vi
List of Figures .....	viii
List of Appendices .....	ix
I. Introduction .....	1
Statement of the Problem .....	4
Purpose of the Study .....	4
Research Questions .....	5
Assumptions .....	5
Delimitations .....	6
Definition of Terms .....	6
II. Review of Literature .....	10
FITNESSGRAM .....	10
Upper-body Strength and Endurance Testing of Children .....	11
Reliability, Validity, and Interrater Reliability Evidence for the FITNESSGRAM Muscular Strength Tests .....	12
Norm- and Criterion-Referenced Standards for Muscular Strength and Endurance Tests .....	16
Criterion-Referenced Agreement .....	18
Judgment, Empirical, and Normative Methods of Setting Criterion- Referenced Standards .....	20

	<b>Survival Analysis .....</b>	<b>23</b>
	<b>Conclusion .....</b>	<b>25</b>
<b>III.</b>	<b>Methods .....</b>	<b>26</b>
	<b>Subjects .....</b>	<b>26</b>
	<b>Tests and Test Administration Procedures .....</b>	<b>26</b>
	<b>Test Administrators .....</b>	<b>28</b>
	<b>Data Collection .....</b>	<b>29</b>
	<b>Analyses .....</b>	<b>31</b>
<b>IV.</b>	<b>Results .....</b>	<b>33</b>
	<b>Description of Participants .....</b>	<b>33</b>
	<b>Criterion-Referenced Agreement .....</b>	<b>36</b>
	<b>Survival Analysis .....</b>	<b>41</b>
	<b>Summary .....</b>	<b>48</b>
<b>V.</b>	<b>Summary and Discussion, Conclusions, and Recommendations .....</b>	<b>49</b>
	<b>Summary and Discussion .....</b>	<b>49</b>
	<b>Conclusions .....</b>	<b>54</b>
	<b>Recommendations .....</b>	<b>57</b>
	<b>References .....</b>	<b>104</b>

## LIST OF TABLES

Table		Page
1.	Validity Coefficients for the Upper-Body Strength Test .....	13
2.	Reliability Coefficients for the Upper-Body Strength Test .....	14
3.	Means and Standard Deviations for Girls' Height, Weight, and the FITNESSGRAM Tests of Upper-body Strength and Endurance Scores (n=182) .....	34
4.	Means and Standard Deviations for Boys' Height, Weight, and the FITNESSGRAM Tests of Upper-body Strength and Endurance Scores (n=201) .....	35
5.	Boys' Percent Agreement Indices Between the Push-Up Test and the FITNESSGRAM' s Alternate Tests of Upper-body Strength and Endurance .....	37
6.	Girls' Percent Agreement Indices Between the Push-Up Test and the FITNESSGRAM' s Alternate Tests of Upper-body Strength and Endurance .....	40
7.	Comparisons of the Survival Curves for 8 through 11-year-old Boys on the FITNESSGRAM Tests of Upper-Body Strength and Endurance .....	42
8.	Comparisons of the Survival Curves for 8 through 11-year-old Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance .....	43

9.	Comparisons of the Survival Curves of Eight-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance ...	44
10.	Comparisons of the Survival Curves of Nine-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance ...	45
11.	Comparisons of the Survival Curves of Ten-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance .....	46
12.	Comparisons of the Survival Curves of Eleven-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance ...	47

## List of Figures

Figure	Page
1. The Flexed-Arm Hang .....	9
2. The Modified Pull-Up .....	9

## LIST OF APPENDICES

Appendix	Page
A. Request Letter and Approval Letter from the Principal of Black Fox Elementary .....	58
B. Request Letter and Approval Letter from the Director of schools .....	61
C. Parent Consent Letter .....	64
D. Institutional Review Board Approval .....	66
E. Oral Script to Students Read by the Physical Education Teacher.....	68
F. FITNESSGRAM Criterion-Referenced Standards for Boys and Girls Five to Seventeen + Years-Old .....	70
G. Contingency Tables by Age and Gender .....	73
H. Survival Curve Performances for Eight and Nine, Nine and Ten, and Ten and Eleven-Year-Old Boys and Girls on the Tests of Strength and Endurance .....	82
I. Survival Curve Performances Between Eight, Nine, Ten, and Eleven-Year-Old Boys and Girls Tests of Strength and Endurance .....	95

## CHAPTER I

### Introduction

In the early 1950s there was concern about children's fitness in the United States. Researchers indicated that European children had higher levels of fitness than children in the United States. The American Association for Health, Physical Education, and Recreation (AAHPER) along with newly formed President's Council on Physical Fitness and Sports developed a national youth fitness testing program using the AAHPER Youth Fitness Test to evaluate the fitness levels of American children. The Youth Fitness Test included performance related tests that measured strength, agility, and endurance along with running and jumping ability (Safrit, 1990 & Morrow, Jackson, Disch, & Mood, 1995).

During the 1970s physical educators and researchers became more interested in the health-related fitness of American children (Safrit, 1990). The goals of health-related testing were to identify the short- and long-term benefits of physical fitness and to ensure that children maintained adequate levels of fitness (Pate & Shephard, 1989). Because the AAHPER Youth Fitness test items such as the 50-yard dash and the standing long jump were not considered health-related fitness items, the American Alliance for Health, Physical Education, Recreation and Dance (AAHPERD) no longer supported the AAHPER Youth Fitness Test and developed the AAHPERD Health-Related Physical Fitness Test.

The emphasis of AIPHERD Health-Related Fitness Tests was measuring health-related traits such as 1) aerobic capacity, 2) flexibility, 3) body composition, and 4) muscular strength and endurance. Tests measuring aerobic capacity, flexibility, body

composition, and upper-body strength and endurance were prevalent in youth fitness test batteries (Pate & Shephard, 1989). Although there have been numerous studies linking mile run scores, skinfold measurement scores, and flexibility scores to adequate health, researchers believe that upper-body muscular strength and endurance are also important components of health-related physical fitness. Adequate upper-body strength is necessary for performing functional and daily activities as well as preventing injury and osteoporosis (Ross & Pate, 1987; Kollath, Safrit, Zhu, & Gao, 1991; Pate, Burgess, Woods, Ross, & Baumgartner, 1993). In addition, physical educators can use muscular fitness test scores to document health-related physical fitness and estimate levels that may yield benefits that carry on into adulthood (Cureton & Warren, 1990; CIAR, 1999). Because of the practicality and the importance of muscular strength and endurance testing, practitioners make valiant efforts to include upper-body strength measures in test batteries (Engleman & Morrow, 1991).

There are numerous fitness test batteries that include measures of upper-body muscular strength and endurance (Physical Best, YMCA Youth Fitness Test, the Chrysler Fund/AAU Test, and the FITNESSGRAM). The FITNESSGRAM health-related physical fitness test battery, developed by the Cooper Institute of Aerobics Research (CIAR), endorsed by AAHPERD, is the latest test battery to contain measures of upper-body strength and endurance. Practitioners have the option of using any one of the following FITNESSGRAM field tests to measure upper-body strength and endurance: (1) the traditional pull-up (2) the modified pull-up (3) the push-up and (4) the flexed-arm hang. In addition, researchers have collected sufficient norm-referenced



reliability and validity evidence for these tests of upper-body strength and endurance (Woods, Burgess, & Pate, 2000).

Children's FITNESSGRAM test scores are interpreted from a criterion-referenced standpoint. Criterion-referenced standards were established in the late 1970s and early 1980s to help indicate levels of physical fitness needed for good health (Cureton & Warren, 1990). The FITNESSGRAM's criterion-referenced standards are used to classify a child as either healthy or unhealthy on a particular health-related trait based on the child's fitness test score (CIAR, 1999). A healthy classification is indicative of a child meeting the FITNESSGRAM criterion-referenced standard established for a particular test item. Individuals who do not meet the standard are classified as unhealthy.

The FITNESSGRAM upper-body strength test standards are different across age groups and between genders (CIAR, 1999). The standards were established by a panel of experts who used a combination of professional judgment, normative data, and empirical data (Cureton & Warren, 1990). Because a practitioner has the option of administering any of the four FITNESSGRAM upper-body strength tests, a child should receive the same healthy/unhealthy classification no matter which test is administered.

Unfortunately there is limited evidence supporting the consistency of classification across tests as well as the suitability of the FITNESSGRAM standards across age groups and genders (Cureton & Warren, 1990; Looney & Plowman, 1990). If the tests are not consistent in classification and the standards are not appropriate, problems can occur when using test scores to classify children as healthy or unhealthy. Misclassification of a child may lead to an inappropriate level of increased participation

in physical activity or a discouragement in participation because the child feels the standard is unachievable. Both outcomes may affect the future of the child's level of fitness (Cureton & Warren, 1990).

If practitioners continue to use the FITNESSGRAM tests of upper body strength and endurance to measure and evaluate children's health, criterion-referenced agreement among the tests must be investigated. In addition, the appropriateness of using different FITNESSGRAM upper-body strength and endurance criterion-referenced standards across age groups and genders should be investigated.

### **Statement of the Problem**

The FITNESSGRAM health-related fitness test battery is the latest test battery that includes tests for measuring children's upper-body strength. To ensure that each test classifies children into the same health category, evidence of criterion-referenced agreements must be established among the tests. Currently, no evidence of percent agreement among the four field tests has been reported. In addition, the current study will compare upper-body strength performances of children across age groups and between genders using survival analysis techniques.

### **Purpose of the Study**

The study has a fourfold purpose:

- 1) To determine the agreement between the alternate tests of upper-body strength and endurance (i.e., the modified pull-up, the flexed-arm hang, and the pull-up) and the FITNESSGRAM recommended push-up test.
- 2) To compare how eight, nine, ten, and eleven-year-old boys perform on each test of upper-body muscular fitness using survival analysis techniques.

- 3) To compare how eight, nine, ten, and eleven-year-old girls perform on each test of upper-body muscular fitness using survival analysis techniques.
- 4) To compare how eight, nine, ten, and eleven-year-old boys and girls on each test of upper-body muscular fitness using survival analysis techniques.

### **Research Questions**

Do the alternate tests of upper-body strength and endurance, (i.e., the modified pull-up, the flexed-arm hang, and the pull-up) produce the same criterion-referenced classification as the FITNESSGRAM recommended push-up test across age groups?

Do eight, nine, ten, and eleven-year-old boys' upper-body muscular fitness survival curves follow the same pattern?

Do eight, nine, ten, and eleven-year-old girls' upper-body muscular fitness survival curves follow the same pattern?

Do eight, nine, ten, and eleven-year-old boys' and girls' survival curves follow the same patterns?

### **Assumptions**

1. Children gave maximal effort on all trials of each test item.
2. The sample size was sufficient for each gender and grade level to calculate the percent agreement for each test item.
3. The sample size was sufficient for each gender and grade level to statistically compare the survival curves.
4. The raters were well trained in test administration.
5. The raters assigned valid scores to each student on each test.

6. The children practiced the test items sufficiently and performance reflected true ability.

### **Delimitations**

1. Only children in grades three through six were selected to participate.
2. Children were recruited via convenience sampling from a Rutherford County elementary school.
3. Only one school from Rutherford County was selected to participate.
4. Measurement of upper-body muscular strength and endurance was delimited to the protocols and test items governed by FITNESSGRAM.

### **Definition of Terms**

Criterion-Referenced Standard. A predetermined standard of performance that indicates whether the child has attained a desired level of performance. The child's performance is compared to the standard rather than to other scores (Baumgartner & Jackson, 1995).

Flexed-arm Hang. A test of muscular strength/endurance. Children raise their bodies off the floor with their arms to a position where the chin is above a chin-up bar, elbows are flexed and the chest is close to the bar (Figure 1). The score is the length of time (in seconds) the position is held (CIAR, 1999).

Healthy. Operationally defined through FITNESSGRAM. Healthy (Healthy Fitness Zone) represents those individuals that scored at or above the set criterion for a test item. This classification is indicative of having some degree of protection against disease that results from a sedentary lifestyle (CIAR, 1999).

**Health-related Physical Fitness (HRPF).** An association with the positive effects of regular, vigorous exercise. The components of HRPF are associated with the prevention of degenerative disease (Baumgartner & Jackson, 1995).

**Modified Pull-up.** A test of muscular strength/endurance. Individuals are positioned on their backs with shoulders placed below a bar. The height of the bar is set one or two inches beyond the grasp of the hands with the palms of the hand facing away. The individuals pull their bodies with the heels placed on the floor, maintaining a straight torso and legs. The individuals complete the exercise by pulling the body up to where the chin meets an elastic band (Figure 2). Individuals are scored on how many successful trials are completed (CIAR, 1999).

**Muscular Endurance.** The ability of a muscle or muscle group to sustain repeated contractions or the ability to apply a constant force for a period of time (Rosato, 1990).

**Muscular Fitness.** Describes the combined status of muscular strength and endurance (American College of Sport Medicine [ACSM], 1995).

**Muscular Strength.** The amount of force a specific muscle or muscle group exerts in one repetition (ACSM, 1995).

**Norm-referenced Standard.** A standard that scores a performance in relation to the performance of other well defined groups on the same test. In other words, childrens' scores are compared with other childrens' scores (Baumgartner & Jackson, 1995).

**Percent Agreement.** The proportion of individuals who receive the same classification on different test items (Mahar, Rowe, Parker, Dawson, & Holt, 1997).

**Reliability.** The degree to which a given test measures the same score or trait over a series of trials, i.e., test-retest (Baumgartner & Jackson, 1995).

**Unhealthy.** Operationally defined through FITNESSGRAM. Unhealthy represents those children who scored below the set criterion for a test item. This classification is indicative of having an increased risk of disease, which results from a sedentary lifestyle (CIAR, 1999).

**Upper-body Strength.** This is defined as the strength of the upper-body muscles such as the latissimus dorsi, pectoralis major, biceps, and triceps. It is operationally defined as the total number of repetitions completed on the modified pull-up test, push-up test, and pull-up test, and the number of seconds completed on the flexed-arm hang test (CIAR, 1999). It is considered an essential part of health-related physical fitness (Pate et. al, 1993; Pate, Ross, Baumgartner, & Sparks, 1987).

**Validity.** The level at which a test item measures what it is supposed to measure, i.e., pull-up and upper-body strength (Baumgartner & Jackson, 1995).

## CHAPTER II

### Review of Literature

The primary purpose of the study was to determine if the four FITNESSGRAM tests of upper-body strength produce the same criterion-referenced classification (i.e., healthy or unhealthy) for children ages eight to eleven. A second purpose of the study was to compare the upper-body strength performances of the children across age groups and by gender, using survival analysis techniques. The purpose of this literature review is to discuss issues that pertain to the purposes of the study. The sections included in this literature review are: (1) FITNESSGRAM, (2) Upper-Body Strength and Endurance Testing of Children, (3) Reliability, Validity, and Interrater Reliability Evidence for the FITNESSGRAM Muscular Strength Tests, (4) Norm-Referenced Versus Criterion-Referenced Scores, (5) Criterion-Referenced Agreement, (6) Judgment, Empirical, and Normative Methods of Setting Criterion-Referenced Standards, (7) Survival Analysis, and 8) Conclusion.

#### **FITNESSGRAM**

During the 1970s, tests such as the AAHPER Youth Fitness test were under scrutiny because they failed to measure health-related aspects of fitness and explain the association between fitness and to day-to-day activities (Pate & Hohn, 1994). In efforts to change fitness testing and measure health-related aspects of fitness, a number of health-related fitness test batteries were developed. In lieu of the AAHPER Youth Fitness test, AAHPERD developed the AAHPERD Physical Best. In later years the YMCA Youth Fitness test, the Chrysler Fund/AAU test and the FITNESSGRAM were developed to measure youth fitness (Pate & Hohn, 1994).

The Cooper Institute for Aerobic Research (CIAR) developed the FITNESSGRAM to measure children's fitness. The FITNESSGRAM is endorsed by AAHPERD, which is a comprehensive health-related fitness assessment and is composed of the following health-related components: 1) body composition, 2) aerobic capacity, 3) muscular strength and endurance, and 4) flexibility. The FITNESSGRAM is unique because it provides the physical educator with a variety of options for each component of fitness. For example, the FITNESSGRAM has four tests available to measure upper-body muscular strength and endurance (i.e., pull-up test, modified pull-up test, flexed-arm hang test and the push-up test). Further, the FITNESSGRAM offers a computerized feedback system, curricular materials, and activity behavior and performance recognition systems (Pate & Hohn, 1994).

### **Upper-body Strength and Endurance Testing of Children**

Muscular strength represents the amount of force a specific muscle or muscle group exerts in one repetition. Muscular endurance is the ability to perform a number of repetitions of a given percentage of maximal weight lifted (American College of Sport Medicine [ACSM], 1995). Muscles exerting force, resisting fatigue, and moving freely through a range of motion with efficiency is indicative of a healthy musculoskeletal system (CIAR, 1999). Good muscular fitness is also necessary for a person to perform daily activities without undue risk of injury (Ross & Pate, 1987).

Historically, children's strength and endurance was measured using four field tests (i.e., pull-up test, modified pull-up test, flexed-arm hang test, and the push-up test). One could find one or more of these tests in test batteries such as the AAHPERD Physical Best, the YMCA Youth Fitness Test, and the Chrysler Fund/AAU test. Unlike



the aforementioned youth fitness test batteries, the FITNESSGRAM recommends the push-up test because a large number of children can be tested at once. Furthermore, the test does not require special equipment and does not produce many zero scores. The modified pull-up, flexed-arm hang, and the pull-up test are optional tests. Practitioners can monitor, document, and provide appropriate activities for children using the scores from these tests (CIAR, 1999).

### **Reliability, Validity, and Interrater Reliability Evidence for the FITNESSGRAM Muscular Strength Tests**

Upper-body strength and endurance testing of children is useful in providing information on children's muscular fitness. Thus, it is important that the tests are free of measurement error (reliability) and are accurate measurements of the trait of interest (validity). Norm-referenced reliability and validity evidence for the four FITNESSGRAM tests of upper-body strength tests are well documented and presented in Tables 1 and 2. The validity and reliability coefficients for muscular strength and endurance range from moderate to high and are deemed acceptable by experts (CIAR, 1999).

Pate, Burgess, Woods, Ross, & Baumgartner (1993) compared field tests (e.g., pull-ups, modified pull-ups, flexed-arm hang, and push-ups) to laboratory tests (e.g., bench press, arm curl, and lat pull-down) among nine- and ten-year-old children. Each child was tested on the three laboratory tests and the next day was administered all field tests. Pate et al., (1993) concluded that the field tests were valid when measuring relative strength (i.e., absolute strength divided by body weight).

Table 1Validity Coefficients for the Upper-Body Strength Tests.

<u>Source</u>	<u>N</u>	<u>Sex</u>	<u>Age</u>	<u>Field/Criterion tests</u>	<u>Validity coefficient</u>
Jackson et al., (1994)	40	M	24.5	90 <sup>0</sup> push-up, bench press	.30
	23	F	24.7	1-RM	.23
Pate et al., (1993)	38	M	9-10	pull-up, lat pull down 1-RM	-.16
	56	F	9-10		.05
		M		push-up, bench press, 1-RM	.36
		F			.02
		M		flexed-arm hang, arm curl 1-RM	.23
		F			.12
Rutherford & Corbin (1993)	204	F		college pull-up, bench press 1-RM	.27
				90 <sup>0</sup> push-up, bench press & lat pull down	.37 & .26
				pull-up, lat pull down 1-RM	.19
				flexed-arm hang, arm curl 1-RM	.26
Woods et al., (2000)	38 56	M F	9-10	pull-up, push-up, modified pull-up, & flexed-arm hang. bench press 1-RM, lat pull 1-RM, & arm curl 1-RM.	.70 - .90**

Note. \*\* Coefficients were higher when accounting for body weight.

Table 2Reliability Coefficients for the Upper-Body Strength Tests.

<u>Source</u>	<u>N</u>	<u>Sex</u>	<u>Grade</u>	<u>Tests</u>	<u>Reliability Coefficients</u> <u>Two Trials - One Trial</u>	
Cotten (1990)	21	M	3	modified pull-up	r = .75	r = .59
	27	F	3		r = .88	r = .78
	33	M	4	r = .90	r = .82	
	37	F	4	r = .92	r = .86	
	31	M	5	r = .79	r = .65	
	33	F	5	r = .83	r = .71	
	29	M	6	r = .90	r = .82	
	33	F	6	r = .95	r = .90	
Engelman & Morrow (1991)	70	M	3	pull-up	r = .95	r = .90
	87	F	3		r = .95	r = .91
	89	M	4		r = .96	r = .92
	74	F	4		r = .95	r = .91
	83	M	5		r = .91	r = .83
	67	F	5		r = .96	r = .92
	242	M	3,4, & 5		r = .94	r = .88
	228	F	3,4,& 5		r = .95	r = .91
McManis et al., (2000)	25	M	3,4,& 5	90° push-up	r = .90	r = .82
	20	F			r = .91	r = .84
	45	M/F	9 & 10	r = .91	r = .83	
	32	M		r = .59	r = .42	
	23	F		r = .94	r = .88	
	55	M/F		r = .75	r = .60	
Pate et al., (1993)	38	M	4 & 5	flexed- arm hang	r = .90	
	56	F			r = .85	

Similar results were indicated in Woods, Burgess, & Pate (2000) among nine- and ten-year-old children. The same findings can be seen in Jackson and Fromme (1994), McManis and Wuest (1994).

Although the pull-up test tends to be a more reliable test of strength among children, it is confounded by body weight and yields too many zero scores (Woods et al., 2000; Kollath, Safrit, Zhu, & Gao, 1991; Pate, Ross, Baumgartner, & Sparks, 1987). Engelman and Morrow (1991) found similarities and differences in administering the traditional pull-up and modified pull-up tests. Engelman & Morrow (1990) found that administering two trials of the traditional pull-up test elicited slightly higher reliability coefficients than the modified pull-up test. In addition, the modified pull-up test did not negate the effects of body composition on upper-body strength performances. Therefore, results from the modified pull-up test were similar to the traditional pull-up test. However, the modified pull-up produced lower percentages of zero scores than the traditional pull-up. Cotten (1990) observed that boys and girls from the National Children Youth Fitness Study II (NCYFS II) were able to perform the modified pull-up with fewer zero scores. Cotten (1990) also observed that the percentage of zero scores for the modified pull-up test among children was lower than the percentage of zero scores associated with the flexed-arm hang and traditional pull-ups reported in other studies. These findings were also consistent with those of Walker and Lloyd (2000) and those of Jackson, Bruya, Baun, Richardson, Weinberg, & Caton (1992). From a practical and motivational standpoint, the modified pull-up may be a better test than the traditional pull-up test. Children can successfully perform the test, thus increasing their likelihood

of achieving behavioral objectives that practitioners set for children as the result of the testing (Engelman & Morrow, 1991; Jackson et al., 1992; & Pate et al., 1987).

Test administrators can affect the reliability of test scores, thus affecting inter-rater reliability or objectivity. Several sources of error can occur in scoring a test. Some of these errors are body position, body motion, and uncorrected performances during each trial. Kollath et al., (1991) investigated the reliability and objectivity of scoring the modified pull-up test where six raters were trained in three major sessions during several weeks of preparation. Inter-rater reliability estimates were .91 for scoring boys and .72 for scoring girls. It was concluded that testers should be extensively trained in administering tests to ensure that each test is administered correctly and scores are recorded accurately. Regardless of which test of upper-body strength and endurance is administered, all four field tests appear to measure upper-body strength and endurance with acceptable levels of reliability and validity when tests are administered by well-trained test administrators (Woods et al., 2000; Pate et al., 1993).

### **Norm- and Criterion-Referenced Standards for Muscular Strength and Endurance Tests**

Traditionally, childrens' fitness test scores have been interpreted from a norm-referenced standpoint; that is, scores are reported as a percentile rank. The normative standard identifies where childrens' performance scores rank relative to the performances of other children tested or a known reference group (Rutherford & Corbin, 1994; Cureton & Warren, 1990). Population-based normative standards can be valuable in interpreting childrens' fitness data. Normative data can also describe the status of a population. When testing takes place periodically, the data can provide a basis for tracking changes

within the observed population. In addition, normative data can be used to compare subgroups within the population at large (Ross, Pate, Delpy, Gold, & Svilar, 1987). However, one of the major concerns with the norm-referenced interpretation of test scores is that a child's rank in a group may not be indicative of the child's true health status on a particular trait.

Criterion-referenced standards were established in the late 1970s and early 1980s to determine what score on a trait is necessary to be free of risk factors that lead to disease (Rutherford & Corbin, 1994; Cureton & Warren, 1990). There are many advantages to using criterion-referenced standards to interpret childrens' fitness test scores. One is that criterion-referenced standards allow the assessment of an attribute that is related to the ability to perform physical activity. Second, the standards are intended to represent the minimal level of a trait that is indicative of good health. One can score very high or very low in relation to the reference group and still meet an acceptable standard. A third advantage of using criterion-referenced standards is that practitioners can provide diagnostic information. A test score can identify a level of physical fitness, which enables the practitioner to decide whether the individual needs to modify activity levels, behavior, or diet. Finally, the fact that a criterion-referenced score is higher or lower than another criterion-referenced score does not mean that the score is better or worse. Both scores can be considered passing or failing depending on whether the scores fall at, above, or below the established criterion-referenced standard. Hence, the purpose of testing and using criterion-referenced standards is to identify individuals as healthy or unhealthy based on the standard, as well as to provide a successful

experience that may lead to positive attitudes towards physical fitness (Pate & Shephard, 1989; Cureton & Warren, 1990; Rutherford & Corbin, 1994; CIAR, 1999).

The FITNESSGRAM interprets childrens' fitness scores using criterion-referenced standards. The criterion-referenced standards are used to classify a child as either healthy or unhealthy on a particular health-related trait based on the child's score on a particular fitness test (CIAR, 1999). Like normative standards, criterion-referenced standards vary according to age and gender. Most professionals acknowledge the fact that boys and girls perform differently because of their maturation. Researchers reported differences in boys' and girls' (grades 5-12) test performances. It was reported that boys performed more sit-ups and chin-ups, stretched farther, and had less body fat as they grew older. Girls' upper-body strength remained consistently low through the early adolescent years. Only abdominal strength and flexibility improved with age among girls. This report contradicts the beliefs that boys' performances on fitness tests tend to peak after puberty and plateau for the remaining years. The report also contradicts the belief that girls' performances peak at the onset of puberty and decline thereafter (Ross & Gilbert, 1985).

### **Criterion-Referenced Agreement**

Because the FITNESSGRAM standards are designed to classify children as either healthy or unhealthy based on their test scores, it is important that students are classified consistently across tests. In other words, regardless of which FITNESSGRAM test of upper-body strength is administered, the child should receive the same criterion-referenced classification. The criterion-referenced agreements among the four

FITNESSGRAM tests of upper-body strength and the appropriateness of the criterion-referenced standards are not well documented (Rutherford & Corbin, 1994).

### Agreement

One method of investigating the criterion-referenced agreement between tests is to use a percent agreement ( $Pa$ ) analysis. A 2 x 2 classification table is used to determine the  $Pa$ .

The formula for percent agreement is:

$$Pa = \frac{A + D}{A + B + C + D}$$

where A represents the number of people classified similarly (healthy) on parallel tests and D represents the number of people classified similarly (unhealthy) on two parallel tests. Values B and C represent people who were not classified the same on the two parallel tests. The sum of A and D is divided by the sum of all consistent and inconsistent classifications for the two parallel tests (Baumgartner & Jackson, 1995). The  $Pa$  coefficient does not consider the possibility of classifications happening by chance. Therefore, a kappa ( $K$ ) coefficient is calculated. The formula for kappa:

$$K = \frac{Pa - Pc}{1 - Pc}$$

$Pc$  equals the proportion of agreement expected by chance. To calculate  $Pc$  all possible combinations of classification for the two parallel tests are divided by the sum of all classifications squared.

$$Pc = \frac{[(A+B)(A+C) + (C+D)(B+D)]}{(A+B+C+D)^2}$$



A meaningful interpretation of  $P_a$  is .50 - 1.00, anything below .50 is deemed unacceptable, whereas kappa coefficients can range from -1.00 to 1.00. A negative value has no meaning in regards to reliability, because test information either contributes to the consistency of classification or it does not. Therefore, a meaningful interpretable range of kappa is .00 to 1.00 (Safrit & Wood, 1995).

Although the criterion-referenced agreement of the FITNESSGRAM upper-body strength tests is not well-documented, Mahar, Rowe, Parker, Mahar, Dawson, & Holt (1997) used percent agreement and modified kappa statistics to assess the parallel forms criterion-referenced reliability of the mile run/walk and PACER test. The two tests purportedly measure aerobic endurance. Children in the fourth and fifth grades were recruited for the study. The children were classified as either pass or fail depending on whether they met the criterion standard on each test. The percent agreement between the two tests was low to moderate for boys and low for girls. Because the agreement among tests was low to moderate for boys and very low for girls, Mahar et al., (1997) concluded that the results may be associated with inappropriate criterion standards and the moderate relationship between the two tests.

### **Judgment, Empirical, and Normative Methods of Setting Criterion-Referenced Standards**

The FITNESSGRAM uses criterion-referenced standards to establish set criterion or cut-off scores for different age groups and gender. The cut-off score or criterion standard represents a desirable health standard achievable by the majority of the population. According to Baumgartner and Jackson (1995), criterion-referenced

standards are based on expert judgment, normative data, and empirical data, or a combination of all three.

The judgmental approach relies on the experience and judgment of experts in the field. The experts decide on what is an acceptable score to be used as a criterion standard on a particular test. The judgmental method is often used and is subject to criticism because of its lack of empirical evidence. The normative approach uses normative population data to establish a cut-off score. Like the judgmental approach, this method lacks validation (Cureton & Warren, 1990).

The empirical approach provides two methods of establishing the cut-off score. First, if there is a criterion measure that relates to the attribute, a cut-off score can be established. Test performances can then be directly associated with the cut-off criterion (Chun, Corbin, & Pangrazi, 2000). The second approach is to use test scores from longitudinal studies of contrasting groups. The scores are distributed and graphed. The cut-off score is established where the two distributions cross. There is still some judgment used in making the decision, but the empirical evidence would help in validating the established criterion (Berk, 1976).

Historically, criterion-referenced standards for muscular fitness have been derived from normative data and professional judgment. Rutherford and Corbin (1994) attempted to establish and validate cut-off scores on the FITNESSGRAM pull-up, push-up and the flexed-arm hang tests in college age women by using the contrasting group method. The contrasting groups consisted of trained and untrained college females. The trained group participated in training the muscles that are utilized during the three tests previously mentioned. The untrained group did not train. The investigators chose this procedure

because there was no known valid criterion measure for muscular strength and endurance. In addition, the investigators felt that they would generate a more valid cut-off score with this procedure.

The contrasting group (trained/untrained) method in the validation and cross validation groups elicited optimal cut-off scores for each test. The optimal cut-off scores yielded misclassification errors ranging from zero to .17 and .06 to .17 for false untrained and false trained classification, respectively. A false untrained classification resulted from a person not passing the standard even after training. A false trained classification represented those who did not train, but passed the standard. Rutherford and Corbin (1994) also examined the passing rates for each test. The passing rates were based on the optimal cut-off score established during the investigations. Passing rates for the untrained group was low to moderate for all three tests. In the trained group the percentages ranged from moderate to high for all three tests. In conclusion, the researchers suggested that a range of scores around the criterion-referenced standard would be applicable and suggested the use of the trained criterion-referenced standard as a benchmark (Rutherford & Corbin, 1994).

Looney and Plowman (1990) noted similar passing rates using the FITNESSGRAM criterion scores. The data from the National Children Youth Fitness Study I and II were used in conjunction with the FITNESSGRAM criterion standards. In both investigations it was recommended that further research with larger samples be done. It was also suggested that applying cut-off scores to both gender and age groups would increase the stability of the criterion-referenced standards (Rutherford & Corbin, 1994; Looney & Plowman 1990).

Looney and Plowman (1990) stated that neither the appropriateness nor the validation of the FITNESSGRAM cut-off scores has been investigated; thus, it is imperative that the appropriateness of these standards be investigated. The standards provide a basis for investigators to evaluate an individual's fitness and health status and to recommend modifications in physical activity, behavior, and diet (Cureton & Warren, 1990).

### **Survival Analysis**

Survival analysis is a collection of statistical procedures for analyzing data in which the outcome variable of interest is time until an event occurs (Kleinbaum, 1995). The event can be any experience that happens to an individual such as death, disease incidence, or failure to continue. Time represents a number in years, months, or days that it takes for the event to happen. Typically, survival analysis is used for epidemiology and clinical studies. The focus of these studies is based on time until an event happens. For example, instead of asking how fast patients change over time, a researcher using survival analysis asks "How much time passes before a change in the client takes place?" In the field of speech pathology, survival analysis can identify a time frame in which one pronounces a word or vowel. Further, survival analysis can compare two methods of speech therapy. Clinicians can determine which therapy will save time and money as well as the effectiveness of the therapy on articulation and vocabulary (Gruber, 1999). Another example of survival analysis is its application in psychotherapy. A clinical psychologist can compare treatments that may hinder reoccurrence of depressive episodes. Two types of therapy, drug therapy versus a placebo, can be examined. The event of interest is failure. Failure in this study represents the time until the patient is

diagnosed with a new depressive condition during the study. Plotting the times of failures in the study, survival curves can identify how each patient, in each group, performed throughout the study. A log-rank test can determine whether the two curves are equivalent (Greenhouse et al., 1989).

One type of survival analysis is a Kaplan-Meier (KM) procedure. Kaplan-Meier is a procedure for plotting and interpreting survival curves. When there is more than one KM survival curve, a log-rank test can be used to identify the statistical equivalence between two or more KM curves (Kleinbaum, 1995). The survival curve represents the cumulative proportion of individuals who have not responded by a fixed point in time. The proportion of individuals who have not responded at the beginning is 100%. As subjects respond, the survival curve decreases. When all individuals respond, the survival curve decreases to zero. Thus, one of the characteristics of the Kaplan-Meier estimate of the survival curve is that the curve decreases only when individuals respond and the decrease is proportional to the number of observed responses. The survival curve ends at the last observed response time (Greenhouse, Stangl, & Bromberg, 1989). Once the survival curves are estimated, a log-rank test can examine the equivalence between two curves. The log-rank test is a large sample chi-square test that compares the overall KM curves. The log-rank test uses observed minus expected counts of failure times for the entire data set (Gruber, 1999). In other words, the log-rank test gives equal weight to each failure time. Each statistical analysis is at  $G-1$  degrees of freedom.  $G$  is the number of survival curves being compared (Kleinbaum, 1995). Development of the formula is quite lengthy and is outlined in three sources: Kleinbaum, (1995), Greenhouse et al., (1989) and Gruber (1999).

Although no studies of survival analysis could be found in the physical education literature, it appears that this analysis is appropriate for analyzing fitness data, especially youth fitness data where children may fail differentially over time on a fitness trait.

Using survival analysis techniques to plot and compare group performances can provide insight on how groups of children perform on a test as well as provide statistical tests for comparing the performance curves of two groups of children. Test developers can use survival analysis to determine whether different criterion-referenced standards are necessary for different groups of children rather than assuming that standards must be increased from year to year.

### **Conclusion**

With continued efforts in testing children for strength, the relationship between levels of muscular strength and endurance in childhood and health status in later adult life can be made clearer. In doing so, the FITNESSGRAM upper-body strength tests that are administered should be reliable and valid. When a variety of FITNESSGRAM upper-body strength tests are available to measure the muscular fitness component, the tests should classify children in a consistent manner. An acceptable level of agreement among the FITNESSGRAM upper-body strength tests can assure that accurate interpretations of those scores are conveyed. In addition, the FITNESSGRAM criterion-referenced standards should be appropriate across age groups and genders. Therefore, the reason for investigating the agreeability and the appropriateness of the criterion-referenced standards of the FITNESSGRAM upper-body strength tests is to enhance the use and understanding of childrens' health-related fitness testing.

## CHAPTER III

### Methods

The primary purpose of the study was to determine if the four FITNESSGRAM tests of upper-body strength elicit the same criterion-referenced classification (i.e., healthy or unhealthy) for children ages eight to eleven. The second purpose of the study was to compare the upper-body strength performances of children across age groups and between genders using survival analysis techniques. The methods used to answer the research questions of interest will be discussed in the following sections: 1) Subjects, 2) Tests and Test Administration Procedures, 3) Test Administrators, 4) Data Collection, and 5) Analyses.

#### **Subjects**

The subjects were a convenience sample of 403 children from an elementary school in Rutherford County. The children were between the ages of seven and thirteen and were enrolled in grades three through six physical education classes. The scores from children ages seven, twelve, and thirteen were not used for the study because of low numbers in these groups. Thus the total number in the sample was 383 (boys n= 201, girls n= 182). Upon permission from the principal and the director of schools as well as a formal consent from the parents, approval from the Institutional Review Board was obtained prior to testing (Appendix A-D).

#### **Tests and Test Administration Procedures**

The FITNESSGRAM (Cooper Institute for Aerobics Research [CIAR], 1999) upper-body strength tests were administered to all participants. The tests included the modified pull-up test, the traditional pull-up test, the flexed-arm hang test, and the 90<sup>0</sup>

push-up test. Test administration procedures were strictly followed as detailed in the FITNESSGRAM test manual, (pp. 25-28).

### Push-up

The children were asked to place the hands shoulder width apart, arms extended, with the legs straight and the toes touching the floor. The back was straight and in line with the head and toes. The children were asked to lower themselves to the point where their elbows were bent at a 90<sup>0</sup> angle and the upper arms were parallel to the floor. Children completed as many push-ups as possible at a pre-determined rhythm (one push-up every three seconds). To ensure the pace, the investigator called out the cadence during test administration. The child's push-up score was the number of correct push-ups completed. The test was stopped for the child when one of the following three things happened: 1) the investigator's verbalization of a second correction, 2) the failure by the child to keep up with cadence, or 3) the failure to achieve a 90<sup>0</sup> with the elbow (CIAR, 1999).

### Modified Pull-up

The modified pull-up required special equipment. A modified pull-up stand was constructed in accordance with the guidelines in the FITNESSGRAM Test Administration Manual, Appendix A (pp. 73). The children were asked to lie on their backs and grasp the bar with an overhand grip. The children started the test in the down position with arms and legs straight. The children were asked to pull their chin to the designated height of the elastic band. The children repeated the movement as many times as possible with only the heels touching the floor. The modified pull-up score was the number of correct pull-ups for each child. The test was stopped for each child after the



investigator verbalized a second correction of form or if the child stopped and rested. The child was also instructed to stop if he/she experienced pain or discomfort (CIAR, 1999).

### Pull-up

The children were asked to assume a hanging position on a bar with an overhand grip; this was the starting position for all children. Assistance was provided for the shorter children to assume the starting position. The children were asked to pull the chin above the bar using only the arms. Kicking and bending of the legs and knees were not permitted. If a child started to swing, the investigator placed an arm in front of the legs to control for excessive swinging. The test was stopped when a second correction was made on form. The child was scored by the number of successful pull-ups (i.e., chin above the bar).

### Flexed-Arm Hang

The children were instructed to grasp the pull-up bar with an overhand grip. The test started as soon as the starting position was established. The starting position was observed at the instant the child's chin was held above the bar with elbows flexed and chest close to the bar. At this time the investigator started the stopwatch. The stopwatch was stopped when the chin dropped below the bar or after the second correction on form was made (e.g., chin touched the bar or when the head tilted back to keep the chin above the bar) was made. The test was scored by the number of seconds the child maintained correct form (CIAR, 1999).

### **Test Administrators**

Test administrators consisted of five graduate assistants and two professors from the Health, Physical Education, Recreation, and Safety Department at Middle Tennessee State University. All test administrators had prior experience administering the selected tests from the FITNESSGRAM test battery. However, to ensure sufficient reliability, validity, and objectivity of the test scores, all test administrators received additional training on each test prior to data collection. Practice trials were completed before the data collection.

### **Data Collection**

The data were collected over a three-week period. During week one, the investigator contacted the physical education teacher. An oral script (Appendix E) was given to the physical education teacher to read to the children about the testing and data collection. In addition, children were instructed on each test item and given a demonstration. After instruction and demonstration the children were given time to practice the FITNESSGRAM strength tests. During this time all raters practiced by following the procedures specified in pages 25-28 of the FITNESSGRAM Test Administration Manual. The practice trials allowed for the identification and remedy of any procedural and or scoring problems.

During week two, the children performed the push-up test, the modified pull-up test, the flexed-arm hang test and the pull-up test, and were measured on height and weight. Because each class period lasted approximately thirty minutes and met on a Monday-Wednesday and Tuesday-Thursday schedule, the testing was completed in two days. On the first day of testing, Monday and Tuesday, the modified pull-up (MPU) and

the flexed-arm hang (FAH) tests were administered. As the children entered the gymnasium they were divided into three groups of eight. Each group started either at the modified pull-up, flexed arm-hang, or the height and weight station.

At the modified pull-up station the test administrator provided instruction and demonstration of the modified pull-up. Each child was given adequate time to perform as many modified pull-ups as possible. The other children waited in line. Upon completion of the modified pull-test the child switched to the next station.

A second administrator recorded shoeless height (in inches) from a wall chart and weight (in pounds) from a stadiometer. To ensure the privacy of the children during the measuring of height and weight the other children were asked to turn their backs to the administrator and child being tested. In addition, all of the children's birth dates were recorded at that station. When all information was completed the child moved to the next station.

At the flexed arm-hang station a third test administrator provided instruction and demonstration prior to testing. Each child was placed into the desired position and was given a verbal command to start. The other children waited in line for their turn.

On the second day of testing (Wednesday-Thursday), the pull-up test (PU), the push-up test (PSU), and an activity station were administered. As the children entered the gymnasium they were divided into three groups of eight. Each group started either at the pull-up, push-up, or activity station.

At the pull-up station a test administrator instructed the children on the testing procedures. Demonstration was also given. Children were asked to stand and wait for their turn. Upon completion the child switched to the next station.

At the activity station children were provided a fun activity while they waited for the next station. The activity consisted of a jump rope station. The jump rope station was not related to the testing. The children had approximately two to three minutes of rest before performing at the next station.

At the push-up station the third and fourth test administrators provided instruction and demonstration prior to testing. Children performed push-ups to the cadence of one push-up every three seconds. To ensure a proper pace an additional administrator called out the cadence. Four children were tested at a time while the others waited for their turn.

On both days of testing, each child stayed approximately eight minutes at each of the three stations. This allowed two to three minutes before a child performed the next test. To ensure that each child received proper rest between all tests, the children were asked to remain in "lunch line order". Lunch line order kept the children in alphabetical order. If a child performed a test first at station one that child also went first at the second and third stations. This procedure was followed during the data collection period.

The third week was allocated for make-up tests and data collection.

## **Analyses**

### **Agreement**

The children were categorized as healthy if they met the criterion-referenced standard for their age and gender and unhealthy if they did not meet the criterion-referenced standard (CIAR, 1999). The healthy, unhealthy criterion-referenced standards for girls and boys across age groups are presented in Appendix F. Percent agreement ( $P_a$ ) was used to evaluate the criterion-referenced parallel test reliability among the four

FITNESSGRAM upper-body strength tests. Kappa and Modified kappa ( $K_q$ ) were used to correct for chance agreement (Baumgartner & Jackson, 1995). All agreement statistics were calculated using SPSS for Windows (v10.0) and the web site <http://www.cpmc.columbia.edu/homepages/chuangj/kappa/> provided statistical support to SPSS and hand calculations of kappa and modified kappa (Chuang, 2001).

### Survival Analysis

Comparison of the children's performance curves on the tests of upper-body strength was performed using survival analysis techniques. The Kaplan-Meier survival analysis technique outlined in Kleinbaum (1995) was used in this investigation. The Mantel-Cox log-rank was used to statistically compare the survival curves. The significance level was set at the .05 level. All survival analyses were done using SPSS for Windows (v10.0).

## CHAPTER IV

### Results

The primary purpose of the study was to determine if the four FITNESSGRAM tests of upper-body strength elicit the same criterion-referenced classification (i.e., healthy or unhealthy) for children ages eight to eleven. A second purpose of the study was to compare the upper-body strength performances of the children across age groups and between genders using survival analysis techniques. The following sections are included in Chapter 4: 1) Description of Participants, 2) Criterion-Referenced Agreement, 3) Survival Analyses, and 4) Summary.

#### **Description of Participants**

Data were collected on 403 children ages seven through thirteen. Data on children ages seven, twelve, and thirteen (n=20) are not reported in the descriptive statistics or used in the criterion-referenced agreement and survival analyses due to the small sample size. The total sample size for analysis purposes was equal to 383 with 201 boys and 182 girls. The age group sample sizes, mean height, mean weight and mean FITNESSGRAM upper-body test scores for girls and boys, ages eight through eleven are reported in Table 3 and Table 4, respectively.

The weight and height of the boys and girls increased with each successive age group. Girls had a large increase in weight from ages nine to eleven, an average of 12.5-lbs. per year. Boys averaged a weight increase of 13-lbs per year from eight to eleven-years-old, with the greatest increase seen from age ten to eleven. Hamill, Drizd, Johnson, Reed, Roche, & Moore, (1979) have reported the national average for height and weight

for boys and girls. In the current study, the majority of the boys and girls were above the national average for height and weight.

**Table 3**

**Means and Standard Deviations for Girls' Height, Weight and the FITNESSGRAM Tests of Upper-body Strength and Endurance Scores (n=182).**

	<u>8</u> n=39	<u>9</u> n=56	<u>10</u> n=44	<u>11</u> n=43
Variable	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Weight	74.4 (24.0)	80.4 (20.2)	92.7 (28.4)	105.4 (36.1)
Height	52.0 (2.2)	53.6 (2.6)	56.1 (2.6)	58.3 (3.3)
MPU	10.6 (6.9)	9.5 (6.8)	8.9 (6.3)	9.9 (5.7)
PSU	5.6 (4.6)	6.5 (6.8)	5.8 (5.5)	7.3 (6.4)
PU	.64 (1.2)	.39 (.99)	.50 (1.1)	.49 (1.5)
FAH	6.0 (7.3)	4.4 (4.4)	6.0 (7.8)	5.6 (8.2)

**Note.** Height is reported in inches and weight is reported in pounds. MPU represents modified pull-up; PSU represents push-up; PU represents pull-up; and FAH represents flexed-arm hang in seconds.

**Table 4**

**Means and Standard Deviations for Boys' Height, Weight and the FITNESSGRAM Tests of Upper-body Strength and Endurance Scores (n=201).**

	<u>8</u> n=46	<u>9</u> n=50	<u>10</u> n=61	<u>11</u> n=44
Variable	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Weight	72.5 (16.9)	81.4 (22.9)	90.0 (26.0)	111.4 (36.8)
Height	52.2 (2.0)	53.8 (2.7)	55.7 (3.1)	59.1 (3.5)
MPU	11.8 (7.2)	12.7 (8.1)	10.6 (6.7)	10.7 (8.5)
PSU	8.4 (5.9)	10.1 (8.1)	9.4 (7.7)	11.1 (8.1)
PU	.89 (1.4)	.96 (1.5)	.95 (1.9)	.86 (1.7)
FAH	7.5 (8.4)	8.9 (9.0)	9.3 (12.5)	5.8 (7.7)

**Note.** Height is reported in inches and weight is reported in pounds. MPU represents modified pull-up; PSU represents push-up; PU represents pull-up; and FAH represents flexed-arm hang in seconds.



### **Criterion-Referenced Agreement**

The push-up test is recommended by the FITNESSGRAM; therefore  $P_a$  indices were computed between the push-up test and all other tests of upper-body strength.  $P_a$  indices between the push-up test and the other tests of upper-body strength, across age groups and between genders are reported in Table 5 and Table 6. The students were categorized as healthy if they met the criterion-referenced standard for their age and gender and unhealthy if they did not meet the criterion-referenced standard (CIAR, 1999). The healthy, unhealthy criterion-referenced standards for girls and boys across age and gender are presented in Appendix F. Kappa and modified kappa ( $K_q$ ) were used to correct for chance agreement. Contingency tables across age and gender for each comparison are reported in Appendix G.

#### **Eight to Eleven-Year-Old Boys**

Based on the Baumgartner and Jackson (1995) guidelines,  $P_a$  indices for eight through eleven-year-old boys were moderate to high (.61 to .86) for all test comparisons. In this study the eight and nine-year-old boys' PSU-MPU and the PSU-FAH  $P_a$  indices were nearly the same and higher than the PSU-PU  $P_a$  index (Table 5). In the ten-year-old group of boys, the  $P_a$  indices were nearly identical for all three comparisons. The  $P_a$  index for PSU-PU (.74) was much higher for ten year-old boys than eight (.61) and nine-year-old boys (.62). The eleven-year-old boys also had high agreement for the PSU-MPU (.86) and PSU-FAH (.75) test comparisons. In addition, agreement for the eleven-year-old boys' PSU-PU (.70) test comparison was slightly lower than ten-year-olds (.74).

Table 5

Boys' Percent Agreement Indices Between the Push-Up Test and the FITNESSGRAM's  
Alternate Tests of Upper-body Strength and Endurance.

Age		PSU-MPU	PSU-FAH	PSU-PU
8 (n=46)	<i>Pa</i>	.78	.80	.61
	K	.39	.56	.28
	<i>Kq</i>	.56	.60	.22
9 (n=50)	<i>Pa</i>	.78	.76	.62
	K	.38	.41	.30
	<i>Kq</i>	.56	.52	.24
10 (n=61)	<i>Pa</i>	.74	.72	.74
	K	.46	.44	.48
	<i>Kq</i>	.48	.44	.48
11 (n=44)	<i>Pa</i>	.86	.75	.70
	K	.70	.53	.46
	<i>Kq</i>	.72	.50	.40

Note. *Pa*= percent agreement. K= kappa. *Kq*= modified kappa.

Typically, boys who were classified as healthy or unhealthy for the PSU test were classified as healthy or unhealthy for the MPU and FAH test, but for the PU test, agreement was lower.

Chuang (2001) reported that a fair to moderate agreement for kappa and modified kappa ranged from .21 to .40 and .41 to .61 respectively. The kappa statistics reported in Table 5 are indicative of a fair to moderate agreement. The kappa statistics for the eight-year-old boys indicated a fair to moderate agreement. The PSU-FAH comparison was the highest among the eight-year-old boys. Nine-year-old boys yielded similar agreement between all three comparisons; all were considered a fair agreement. Like the nine-year-olds, the ten-year-old boys' kappa statistics for the three test comparisons were similar but slightly higher ranging from .44 to .48, indicating moderate agreement. The eleven-year-old boys' kappa statistics ranged from moderate to substantial agreement, with the PSU-MPU comparison being the highest at .70.

Modified Kappa ( $kq$ ) indices for all three test comparisons across the age groups were also fair to moderate. For the eight-year-old boys in the study, the PSU-FAH comparison yielded the highest  $kq$  with the PSU-PU comparison yielding an extremely low  $kq$  (.22). The PSU-MPU and PSU-FAH comparisons were similar for the nine-year-old boys at .56 and .52 respectively. The ten-year-old boys reported the lowest  $kq$  for all three test comparisons. Finally, the PSU-MPU comparison for the eleven-year-old boys had the highest  $kq$  overall at .72 (Table 5).

### Eight to Eleven-Year-Old Girls

The girls'  $P_a$  indices were quite different from the boys' across test comparisons. Unlike boys, the  $P_a$ s were higher for the PSU-PU comparison and lower for the PSU-MPU and PSU-FAH comparisons for eight, nine, and ten-year-old girls (Table 6). The PSU-MPU comparisons were the lowest for all ages.

The eight-year-old girls'  $P_a$  indices for the PSU-MPU and PSU-FAH comparisons were similar at .54 and .56 respectively, while the PSU-PU agreement was higher at .72. For the nine-year-old girls the PSU-PU agreement was the highest at .77, with the PSU-FAH being slightly lower at .64. The PSU-MPU comparison was the lowest at .48. Ten-year-old girls'  $P_a$  indices was the highest for the PSU-FAH and PSU-PU comparisons at .75 and .82 respectively.  $P_a$  indices for the eleven-year-old girls were the same at .67 for the PSU-FAH and PSU-PU comparisons and lower than the ten-year-old girls. In general, girls had lower  $P_a$  indices across the three test comparisons than the boys.

The girls' kappa statistics were also lower than the boys' kappa statistics ranging from poor to moderate agreements. Eight-year-old girls' kappa was the same for the PSU-MPU and PSU-FAH comparisons at .13. The PSU-PU comparison yielded the highest kappa at .41. For nine-year-old girls the PSU-MPU yielded a slight agreement of .09 whereas the PSU-FAH and PSU-PU comparisons were higher at .31 and .48 respectively indicating a fair agreement. A similar pattern is seen in ten and eleven-year-old girls with ten year-old girls' PSU-FAH and PSU-PU comparisons being the highest at .49 and .55 respectively.

**Table 6**

**Girls' Percent Agreement Indices Between the Push-Up Test and the FITNESSGRAM's  
Alternate Tests of Upper-body Strength and Endurance.**

<b>Age</b>	<b>Statistic</b>	<b>PSU-MPU</b>	<b>PSU-FAH</b>	<b>PSU-PU</b>
8 (n=39)	<i>Pa</i>	.54	.56	.72
	K	.13	.13	.41
	<i>Kq</i>	.08	.12	.44
9 (n=56)	<i>Pa</i>	.48	.64	.77
	K	.09	.31	.48
	<i>Kq</i>	-.04	.28	.54
10 (n=44)	<i>Pa</i>	.59	.75	.82
	K	.29	.49	.55
	<i>Kq</i>	.18	.50	.64
11 (n=43)	<i>Pa</i>	.58	.67	.67
	K	.18	.34	.34
	<i>Kq</i>	.16	.34	.34

Modified kappa statistics were also lower than the boys ranging from poor to moderate (Table 6). The PSU-PU comparison for all age groups yielded the highest  $kq$ . The PSU-MPU comparison had the lowest  $kq$  for eight, nine, ten, and eleven-year-old girls as well. The ten-year-old girls' PSU-FAH and PSU-PU  $kq$  were different at .50 and .64 respectively. The eleven-year-old girls' PSU-FAH and PSU-PU  $kq$  were similar with both being at .34.

### **Survival Analysis**

A second purpose of the study was to compare children's performance curves across age and gender for each test of upper-body strength using survival analysis techniques. The Mantel-Cox log-rank was used to statistically compare survival curves. The significance value was set at the .05 level. The survival curves for eight and nine, nine and ten, and ten and eleven-year-old boys and girls on each test of upper-body strength and endurance are represented in Appendix H, Figures 3 through 14 and Figures 15 through 26 respectively.

#### **Eight to Eleven-Year-Old Boys and Girls**

The mean survival time, the standard error, the ninety-five percent confidence interval (95%CI), the Mantel-Cox Log-rank test statistics, and the p values are reported in Tables 7 & 8 for eight to eleven-year-old boys and girls respectively.

There were no significant differences between eight and nine, nine and ten, and ten and eleven-year-old boys' and girls' survival curves on the FITNESSGRAM test of upper-body strength and endurance.

**Table 7**

**Comparison of the Survival Curves for 8 through 11-year-old Boys on the FITNESSGRAM Tests of Upper-Body Strength and Endurance.**

Age	Test	Mean(95%CI)	SE	Log Rank	p value
8 (n=46)	MPU	11.8± 2.1	1.1	.42 <sup>a</sup>	.52
	PSU	8.4± 1.7	.87	1.71 <sup>a</sup>	.19
	PU	.89± .39	.20	.15 <sup>a</sup>	.70
	FAH	7.5± 1.2	1.2	.53 <sup>a</sup>	.47
9 (n=50)	MPU	12.7± 2.3	1.1	1.82 <sup>b</sup>	.18
	PSU	10.1± 2.2	1.1	.25 <sup>b</sup>	.62
	PU	.96± .42	.21	.00 <sup>b</sup>	.96
	FAH	8.9± 2.5	1.3	.01 <sup>b</sup>	.91
10 (n=61)	MPU	10.6± 1.7	.86	.01 <sup>c</sup>	.92
	PSU	9.4± 1.9	.98	.93 <sup>c</sup>	.34
	PU	.95± .48	.24	.08 <sup>c</sup>	.78
	FAH	9.3± 2.3	1.6	2.97 <sup>c</sup>	.09
11 (n=44)	MPU	10.7± 2.5	1.3		
	PSU	11.1± 4.3	1.2		
	PU	.86± .50	.26		
	FAH	5.8± 2.3	1.2		

**Note.** <sup>a</sup> comparison of 8 & 9-year-olds. <sup>b</sup> comparison of 9 & 10-year-olds. <sup>c</sup> comparison of 10 & 11-year-olds.

**Table 8**

**Comparison of the Survival Curves for 8 through 11-year-old Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance.**

Age	Test	Mean(95%CI)	SE	Log Rank	p value
8 (n=39)	MPU	10.6± 2.2	1.1	.81 <sup>a</sup>	.37
	PSU	5.6± 1.5	.74	.38 <sup>a</sup>	.54
	PU	.64± .38	.20	.78 <sup>a</sup>	.38
	FAH	6.0± 2.3	1.2	1.83 <sup>a</sup>	.18
9 (n=56)	MPU	9.5± 1.8	.91	.03 <sup>b</sup>	.87
	PSU	6.5± 1.8	.90	.35 <sup>b</sup>	.56
	PU	.39± .26	.13	.31 <sup>b</sup>	.58
	FAH	4.4± 1.1	.59	1.32 <sup>b</sup>	.25
10 (n=44)	MPU	8.9± 1.9	.95	.35 <sup>c</sup>	.56
	PSU	5.8± 1.6	.84	1.08 <sup>c</sup>	.30
	PU	.50± .33	.17	.01 <sup>c</sup>	.93
	FAH	6.0± 2.3	1.2	.10 <sup>c</sup>	.76
11 (n=43)	MPU	9.9± 1.7	.87		
	PSU	7.3± 1.9	.97		
	PU	.49± .46	.23		
	FAH	5.6± 2.4	1.3		

**Note.** <sup>a</sup> comparison of 8 & 9-year-olds. <sup>b</sup> comparison of 9 & 10-year-olds. <sup>c</sup> comparison of 10 & 11-year-olds.



### Eight-Year-Old Boys and Girls

The survival curves between eight, nine, ten, and eleven-year-old boys and girls on each test of upper-body strength and endurance are represented in Appendix I, Figures 27 through 42.

Comparisons of eight-year-old boys' and girls' survival curves on the FITNESSGRAM tests of upper-body strength and endurance are reported in Table 9. There was a statistically significant difference ( $p < .05$ ) between eight-year-old boys' and girls' survival curves on the push-up test. All other survival curve comparisons were not statistically different ( $p > .05$ ).

Table 9

Comparison of the Survival Curves of Eight-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance.

Test	Boys (n=46)	Girls (n=39)	Log Rank	p value
	Mean (SE)	Mean (SE)		
MPU	11.8± 1.1	10.6± 2.2	.50	.48
PSU	8.4± .87	5.6± 1.5	5.92*	.02*
PU	.89± .20	.64± .38	.60	.44
FAH	7.5± 1.2	6.0± 2.3	.54	.46

Note. \* indicates a significant difference ( $p < .05$ ).

### Nine-Year-Old Boys and Girls

Comparisons of nine-year-old boys' and girls' survival curves on the FITNESSGRAM tests of upper-body strength and endurance are reported in Table 10. There was a statistically significant difference ( $p < .05$ ) between nine-year-old boys' and girls' survival curves on all four tests of muscular strength and endurance.

Table 10

Comparison of the Survival Curves of Nine-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance.

Test	Boys (n=50)	Girls (n=56)	Log Rank	p value
	Mean (SE)	Mean (SE)		
MPU	12.7± 1.1	9.5± .91	5.19*	.02*
PSU	10.1± 1.1	6.5± .90	4.29*	.02*
PU	.96± .21	.39±.26	5.33*	.04*
FAH	8.9± 1.3	4.4± .59	11.26*	.00*

Note. \* indicates a significant difference ( $p < .05$ ).

Ten-Year-Old Boys and Girls

Comparisons of ten-year-old boys' and girls' survival curves on the FITNESSGRAM tests of upper-body strength and endurance are reported in Table 11. There was a statistically significant difference ( $p < .05$ ) between ten-year-old boys' and girls' survival curves on the push-up test. All other survival curve comparisons were not statistically different ( $p > .05$ ).

Table 11

Comparison of the Survival Curves of Ten-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance.

Test	Boys (n=61)	Girls (n=44)	Log Rank	p value
	Mean (SE)	Mean (SE)		
MPU	10.6± .86	8.9± .95	1.42	.23
PSU	9.4± .98	5.8± .84	7.78*	.01*
PU	.95± .24	.50± .17	2.30	.13
FAH	9.3± 1.6	6.0± 1.2	2.27	.13

Note. \* indicates a significant difference ( $p < .05$ ).

### Eleven-Year-Old Boys and Girls

Comparisons of eleven-year-old boys' and girls' survival curves on the FITNESSGRAM tests of upper-body strength and endurance are reported in Table 12. There was a statistically significant difference ( $p < .05$ ) between eleven-year-old boys' and girls' survival curves on the push-up test. All other survival curve comparisons were not statistically different ( $p > .05$ ).

Table 12

Comparison of the Survival Curves of Eleven-Year-Old Boys and Girls on the FITNESSGRAM Tests of Upper-Body Strength and Endurance.

Test	Boys (n=44)	Girls (n=43)	Log Rank	p value
	Mean (SE)	Mean (SE)		
MPU	10.7± 1.3	9.9± .87	.58	.45
PSU	11.1± 1.2	7.3± .97	5.34*	.02*
PU	.86± .26	.49± .23	1.32	.25
FAH	5.8± 1.2	5.6± 1.3	.00	.96

Note. \* indicates a significant difference ( $p < .05$ ).

## Summary

The four FITNESSGRAM tests of upper-body strength and endurance ranged from poor to moderate in classifying children into healthy/unhealthy categories except for the PSU-MPU comparison for nine-year-old girls which yielded a unacceptable  $P_a$  index of .48. There were higher agreements between the PSU-MPU and PSU-FAH comparisons for eight to eleven-year-old boys and between the PSU-PU and PSU-FAH comparisons for eight to ten-year-old girls (Tables 5 & 6).

As for the survival curves, there were no statistically significant differences between eight and nine, nine and ten, and ten and eleven-year-old boys' and girls' strength and endurance performances on the MPU, PSU, FAH, and PU tests (Tables 7 & 8). However, within gender groups there were statistically significant differences between eight, nine, ten, and eleven-year-old boys' and girls' survival curves on the push-up test. This was also true for nine-year-old boys' and girls' survival curves on the modified pull-up test, flexed-arm hang, and pull-up test (Tables 9-12).

## CHAPTER V

### Summary and Discussion, Conclusions, and Recommendations

The primary purpose of the study was to determine if the four FITNESSGRAM tests of upper-body strength and endurance produce the same criterion-referenced classification (i.e., healthy or unhealthy) for children ages eight to eleven. A second purpose of the study was to compare the upper-body strength performances of the children across age levels and by gender using survival analysis techniques. The following sections are included in Chapter five: 1) Summary and Discussion 2) Conclusions, and 3) Recommendations.

#### **Summary and Discussion**

##### Percent Agreement ( $P_a$ ), Kappa, and Modified Kappa

Research question one dealt with the criterion-referenced agreement between the FITNESSGRAM recommended test of upper-body strength and endurance (i.e., push-up test) and the FITNESSGRAM alternative tests of upper-body strength and endurance (i.e., modified pull-up test, flexed-arm hang test, and pull-up test). The  $P_a$  indices for the PSU-MPU and PSU-FAH test comparisons for eight to eleven-year-old boys ranged from .72 to .86 indicating a moderate to high agreement (Table 5). However, there was only one  $P_a$  index greater than .80 and that was the PSU-MPU comparison with a  $P_a$  index of .86 for eleven-year-old boys. The other  $P_a$  values were between .72 and .80. Overall, the PSU-MPU  $P_a$  indices were the highest across all age groups with  $P_a$  values ranging from .74 to .86 and the PSU-PU  $P_a$  indices were the lowest across all age groups with  $P_a$  values ranging from .61 to .74.

Because *Pa* values may provide an overly optimistic estimate of agreement, kappa and modified kappa statistics were computed to take into account chance agreement. Kappa and modified kappa statistics across all three test comparisons ranged from .28 to .56 and .22 to .60 respectively for eight-year-old boys. Based on Chuang's (2001) criteria, *Pa* indices were interpreted as slight to moderate with most indices being above .41 (moderate agreement). The nine-year-old boys' kappa and modified kappa statistics for the three test comparisons were similar to the eight-year-old boys. The kappa and modified kappa statistics for the nine-year-old boys ranged from .30 to .41 and .24 to .56 respectively, indicating a slight to moderate agreement. The ten-year-old boys' kappa and modified kappa statistics for the three test comparisons ranged from .44 to .48 respectively, indicating a moderate agreement. Eleven-year-old boys' kappa and modified kappa statistics yielded the highest agreements ranging from .46 to .70 and .40 to .72 respectively, indicating a moderate to a substantial agreement (Chuang, 2001).

Girls' *Pa* indices were quite different from the boys' *Pa* indices. There was a moderate agreement for the boys' PSU-MPU and PSU-FAH *Pa* indices. However, the girls' *Pa* indices for the same test were .48 to .75 indicating a poor to moderate agreement. On the other hand, girls had a better PSU-PU *Pa* index than boys did with ten-year-old girls yielding the highest agreement at .82. In fact, girls as a whole had higher *Pa* indices with the PSU-PU comparison than boys did at .67 to .82 and .61 to .74 respectively. When kappa and modified kappa were calculated to account for chance agreement, girls' agreements were quite lower than the boys' agreements. Kappa and modified kappa statistics across all three test comparisons ranged from .13 to .41

and .08 to .44 respectively for eight-year-old girls. Based on Chuang's (2001) criteria, these  $P_a$  values should be interpreted as slight to fair. The nine-year-old girls' kappa and modified kappa statistics for the three test comparisons were similar to the eight-year-old girls. The kappa and modified kappa statistics for the nine-year-old girls ranged from .09 to .48 and -.04 to .54 respectively, indicating a slight to moderate agreement. The ten-year-old girls' kappa and modified kappa statistics for the three test comparisons ranged from .29 to .55 and .18 to .64 respectively, indicating a slight to substantial agreement. The PSU-PU modified kappa index was the highest at .64. Eleven-year-old girls' kappa and modified kappa statistics ranged from .18 to .34 and .16 to .34 respectively, indicating a slight to fair agreement (Chuang, 2001).

In summary, the agreement among the FITNESSGRAM four tests of upper-body strength and endurance ranged from moderate to high for boys eight to eleven-years-old with most agreement indices being moderate. There were only two comparisons with a  $P_a$  index of .80 or higher. The two comparisons with high agreement were the PSU-FAH comparison for eight-year-old boys and the PSU-MPU comparison for eleven-year-old boys. Overall, moderate agreements were found between the PSU-MPU and the PSU-FAH comparisons. When chance agreement was accounted for the agreements were fair to substantial. Among eight to eleven-year-old girls, the  $P_a$  indices were poor to moderate. Only the PSU-PU comparison was above .80 for ten-year-old girls. Overall, low to moderate agreements were found between the PSU-PU and PSU-FAH comparisons. When chance agreement was accounted for the agreement indices were poor to moderate.



In this investigation, there was a lack of substantial agreement between the FITNESSGRAM alternative tests of upper-body strength and endurance and the FITNESSGRAM recommended test of upper-body strength and endurance, the push-up test. The lack of agreement between the push-up test and alternative tests could be due to too stringent or too lenient criterion-referenced standards on some or all of the tests. Another reason for the lack of agreement could be associated with the fact that the four tests of upper-body strength and endurance require the use of different muscle groups. For example, the push-up test measures strength and endurance of the pectoralis major and triceps whereas the pull-up test measures the strength and endurance of the latissimus dorsi and biceps.

Because the alternative tests lack substantial agreement with the FITNESSGRAM recommended test (the push-up test), practitioners and test developers should take notice. First, practitioners should use caution when deciding to use the alternative tests of upper-body strength and endurance. In this investigation, a high percentage of children received a different unhealthy/healthy classification depending upon which test of upper-body strength and endurance was administered. The test developers should take a closer look at the criterion-referenced standards set for each test. It is possible that the criterion-referenced standards need to be altered to maximize test agreement. Test developers may also want to consider whether upper-body strength and endurance should be measured by multiple tests because upper-body strength and endurance is possibly a multidimensional trait.

### Survival Analysis

Research question two dealt with comparing the FITNESSGRAM test performances of eight and nine-year-old boys, nine and ten-year-old boys, and ten and eleven-year-old boys using survival analysis techniques. There were no statistically significant differences between the FITNESSGRAM test performances of eight and nine, nine and ten, and ten and eleven-year-old boys on the modified pull-up test, pull-up test, flexed-arm hang test, and the push-up test (Table 7 & Figures 3 –14). Based on the survival analysis results, performance on the tests of upper-body strength and endurance may not change from age group to subsequent age group for boys eight to eleven years of age. However, it should be noted that longitudinal studies comparing survival curves over time need to be done to evaluate change or lack of change.

Research question three dealt with comparing the FITNESSGRAM test performances of eight and nine-year-old girls, nine and ten-year-old girls, and ten and eleven-year-old girls using survival analysis techniques. There were no statistically significant differences between the FITNESSGRAM test performances of girls ages eight and nine, nine and ten, and ten and eleven on the modified pull-up test, pull-up test, flexed-arm hang test, and the push-up test (Table 8 & Figures 15 –26). As with eight to eleven-year-old boys, the girls' performances did not differ from age group to subsequent age group. Whether this pattern would hold in a longitudinal study needs investigation.

Research question four dealt with comparing the FITNESSGRAM test performances of eight, nine, ten, and eleven-year-old boys and girls. There were statistically significant differences between the boys' and girls' survival curves on the

push-up test across all age groups. There were also statistically significant differences between nine-year-old boys' and girls' survival curves on the modified pull-up test, pull-up test, and flexed-arm hang test.

In summary, survival analysis results did not support the findings of Pate and Shephard (1989) and Malina and Bouchard (1991) who indicated boys' and girls' strength and endurance levels improve with increasing age. However, survival analysis did support Pate and Shephard's (1989) and Malina and Bouchard's (1991) conclusions that strength and endurance levels between boys and girls are different, with the boys' strength and endurance levels being higher. Those strength differences only held up for the push-up test across all age groups. It is possible that the sample sizes used in this investigation may not have been large enough to detect any small performance differences between age groups. However, visual inspection of the survival curves revealed no differences in performance from age group to subsequent age group for boys or girls. It is also possible that this was a unique sample of children and the failure to find year to year performance differences would not hold up in other geographical areas.

## **Conclusions**

Research question 1:

Do the alternate tests of upper-body strength and endurance, (i.e., the modified pull-up, the flexed-arm hang, and the pull-up) produce the same criterion-referenced classification as the FITNESSGRAM recommended push-up test across age groups?

Based upon the characteristics of the sample and the limitations of the study, the following conclusions seem warranted.

1. Based on the large number of poor to moderate agreement indices, using the FITNESSGRAM alternative tests of upper-body strength and endurance will result in different healthy/unhealthy classifications for a high percentage of children, especially girls.
2. The modified pull-up test and the flexed-arm hang test has higher agreement indices with the push test when classifying eight to eleven-year-old boys as healthy/unhealthy than eight to eleven-year-old girls.
3. The pull-up test has higher agreement indices with the push test when classifying eight to eleven-year-old girls as healthy/unhealthy than eight to eleven-year-old boys.
4. Based on the wide range and the poor to moderate agreements for all test comparisons, the four tests may measure different components of upper-body strength and endurance.

Research question 2:

Do eight, nine, ten, and eleven-year-old boys' upper-body muscular fitness survival curves follow the same pattern?

Based upon the characteristics of the sample and the limitations of the study, the following conclusions seem warranted.

1. The four tests of upper-body strength and endurance do not differentiate the strength differences that are typically seen in boys from age group to subsequent age group.

2. The boys' strength changes may have not been seen from age group to subsequent age group because the children were not tested in their previous year.
3. A larger sample size may be needed to detect the differences in boys' strength and endurance levels from year to year.

Research question 3:

Do eight, nine, ten, and eleven-year-old girls' upper-body muscular fitness survival curves follow the same pattern?

Based upon the characteristics of the sample and the limitations of the study, the following conclusions seem warranted.

1. The four tests of upper-body strength and endurance do not differentiate the strength differences that are typically seen in girls from age group to subsequent age group.
2. The girls' strength changes may have not been seen from age group to subsequent age group because the children were not tested in their previous year.
3. A larger sample size may be needed to detect the differences in girls' strength and endurance levels from year to year.

Research question 4:

Do eight, nine, ten, and eleven-year-old boys and girls survival curves follow the same pattern?

Based upon the characteristics of the sample and the limitations of the study, the following conclusions seem warranted.

1. The push-up test identifies that eight to eleven-year-old boys' strength and endurance is higher than eight to eleven-year-old girls.
2. The modified pull-up test, the flexed-arm hang test, and the pull-up test identifies that nine-year-old boys' strength and endurance is higher than nine-year-old girls.

### **Recommendations**

As a result of this investigation, the following recommendations seem appropriate.

1. Replicate the study using a large number of children from different geographical areas.
2. Physical educators should administer two different tests to accommodate the differences in the muscles groups used to perform each test and to understand and justify reasons for classifying children as healthy/unhealthy.
3. Use the current data set to analyze all possible pair-wise comparisons to determine which upper-body strength and endurance tests have the highest criterion-referenced agreement.
4. Alter the criterion-referenced standard on the FITNESSGRAM alternative tests of upper-body strength and endurance to maximize agreement with the push-up test.
5. Design a longitudinal study to compare survival curves across time to assess changes in children's muscular fitness performances.

**Appendix A**

**Request Letter and Approval Letter from the Principal of Black Fox Elementary**

September 13, 2000

Mr. Zane Cantrell, Principal  
Black Fox Elementary  
Murfreesboro, TN 37130

Dear Mr. Cantrell,

I am a doctoral student at Middle Tennessee State University and am conducting dissertation research on measurement issues regarding the Prudential FITNESSGRAM muscular fitness tests. I will like to test students at Black Fox Elementary in order to determine the agreeability between muscular fitness tests as well as examine criterion scores among gender and grade levels. The physical education teacher at Black Fox Elementary currently uses the test battery protocol.

All test results will be confidential. I will only use the test scores to perform the analysis. Once all data were collected and analyzed, all names will be destroyed. The children will benefit directly from testing by receiving feedback on healthy and unhealthy performance levels. Further, the data analyses can be beneficial to the teacher allowing the determination of the agreeability among the four muscular fitness tests. The physical educator can then decide which tests are appropriate for gender and grade levels in determining healthy and unhealthy levels of fitness.

Since testing is already part of the physical education curriculum, the code of Federal Regulations (Title 34 – Education, Part 97 – Protection of Human Subjects) does not mandate that student or parent permission be obtained. However, I will provide notification (enclosed) before data collection. Additionally, I will seek student assent (enclosed) via Mr. Mike Vaughn. I am also seeking consent from the Director of Schools. If you will allow me to collect data, please sign the enclosed letter of approval.

I thank you for your time and consideration. Please feel free to contact me with any questions, 898-5545.

Thank you,

Todd Sherman  
Doctor of Arts Candidate  
MTSU



**Letter of Approval – Principal**

Department of Health, Physical Education, Recreation, and Safety

Middle Tennessee State University

---

  
Principal Investigator

---

  
Responsible Faculty Member

**Project Title: Criterion-referenced agreement of the FITNESSGRAM Upper Body Tests of Muscular Strength and Endurance.**

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 22, 2000.

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

---

  
Signature of Principal

---

  
Date

**Appendix B**

**Request Letter and Approval Letter from the Director of Schools**

September 13, 2000

Ms. Marilyn Mathis, Director of Schools  
Murfreesboro City Schools  
Murfreesboro, TN 37130

Dear Ms. Mathis,

I am a doctoral student at Middle Tennessee State University and am conducting dissertation research on measurement issues regarding the Prudential FITNESSGRAM muscular fitness tests. I will like to test students at Black Fox Elementary in order to determine the agreeability of the four muscular fitness tests. In addition, I am examining the appropriateness of the criterion scores among gender and grade levels. Currently the physical education teacher already uses the test battery protocols I wish to administer. All information collected will be confidential. Upon collection and analysis of data, all names will be purged from my files.

As a result of testing, the children will benefit by receiving feedback on healthy and unhealthy performance levels. Further, the data analyses can be beneficial to the teacher by allowing the determination of the agreeability among the four muscular fitness test. The physical educator can then decide which tests are appropriate for gender and grade levels in determining healthy and unhealthy levels of fitness.

Since testing is already part of the physical education curriculum, the code of Federal Regulations (Title 34 – Education, Part 97 – Protection of Human Subjects) does not mandate that student or parent permission be obtained. However, I will provide notification (enclosed) before data collection. Additionally I will seek student assent (enclosed) via Mr. Mike Vaughn. If a parent contacts the physical education teacher or me I will not use his/her child's test score. I have already received consent from the physical education teacher at Black Fox Elementary. I am also seeking consent from the principal at Black Fox Elementary (Mr. Zane Cantrell).

If you will allow me to collect scores, please sign the enclosed letter of approval.

I thank you for your time and consideration. Please feel free to contact me with any questions, 898-5545.

Thank you,

Todd Sherman  
Doctor of Arts Candidate  
MTSU

**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: Criterion-referenced agreement of the FITNESSGRAM Upper Body Tests of Muscular Strength and Endurance.**

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 22, 2000.

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

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

\_\_\_\_\_  
Date

**Appendix C**

**Parent Consent Letter**

Dear Parent(s)/Guardian(s),

Your child will be participating in fitness testing during the week of **October 30, 2000.**

Fitness tests allow your child to understand if he/she is in good physical shape. I will be assisting Mike Vaughn, the physical education teacher. Further, I will be recording your child's scores for research purposes. The research will investigate how accurate the tests classify your child among four muscular fitness tests: (1) traditional pull-up (2) modified pull-up (3) push-up (4) flexed arm-hang.

All names will be kept confidential. Once all data have been collected, student names will be deleted from the data file. No one will have access to your child's scores except two assistants, the physical educator (Mike Vaughn), and myself. Your child's grade will not be affected by your decision. **If you agree to allow me to record your child's score, please sign at the bottom and return to the physical education teacher by October 27, 2000.** Please understand that your child will be allowed to participate in the study unless you deny permission.

Thank you,

Todd Sherman  
Doctor of Arts Candidate  
MTSU

**\*Sign below if you agree to allow the researcher to record your child's score.**

Parent's signature	Child's Name	Date

**Appendix D**

**Institutional Review Board Approval**

**Elementary and Special Education Department**

P.O. Box 69  
Middle Tennessee State University  
Murfreesboro, Tennessee 37132  
(615) 898-2680

To: Todd Sherman

From: Nancy Bertrand, Chair *Nancy Bertrand*  
MTSU Institutional Review Board

Re: "Criterion-Referenced Agreement of the FITNESSGRAM Upper  
Body Tests of Muscular Fitness"  
Protocol # 01-018

Date: October 23, 2000

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. Dianne Bartley

A Tennessee Board of Regents Institution

*MTSU is an equal opportunity, non-racially identifiable, educational institution that does not discriminate against individuals with disabilities.*





## Appendix E

### Oral script to Students Read by the Physical Education Teacher

Next week we will start fitness testing. This year I will have a helper, his name is Mr. Todd Sherman. Mr. Sherman will also like to record your test scores to determine how well the tests work. If you do not want Mr. Sherman 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 Mr. Sherman to record your scores.

**Appendix F****FITNESSGRAM Criterion-Referenced Standards for Boys and Girls****Five to Seventeen + Years-Old**

## BOYS

Age	One-mile run minutes		PACER # laps		Walk test & VO <sub>2</sub> max ml/kg/min		Percent fat		Body mass index		Curl-up # complete	
5							25	10	20	14.7	2	10
6	Completion of distance. Time standards not recommended.		Participation in run. Lap count standards not recommended.				25	10	20	14.7	2	10
7							25	10	20	14.8	4	14
8							25	10	20	15.1	6	20
9							25	10	20	15.2	9	24
10	11:30	9:00	23	61	42	52	25	10	21	15.9	12	24
11	11:00	8:30	23	72	42	52	25	10	21	15.8	15	28
12	10:30	8:00	32	72	42	52	25	10	22	16.0	18	36
13	10:00	7:30	41	72	42	52	25	10	23	16.8	21	40
14	9:30	7:00	41	83	42	52	25	10	24.5	17.5	24	45
15	9:00	7:00	51	94	42	52	25	10	25	18.1	24	47
16	8:30	7:00	61	94	42	52	25	10	26.5	18.5	24	47
17	8:30	7:00	61	94	42	52	25	10	27	18.8	24	47
17+	8:30	7:00	61	94	42	52	25	10	27.6	19.0	24	47

Age	Trunk lift inches		Push-up # complete		Modified pull-up # complete		Pull-up # complete		Flexed arm hang seconds		Back-saver sit & reach** inches		Shoulder stretch
5	6	12	3	8	2	7	1	2	2	6		8	
6	6	12	3	8	2	7	1	2	2	8		8	
7	6	12	4	10	3	9	1	2	3	8		8	
8	6	12	5	13	4	11	1	2	3	8		8	
9	6	12	6	15	5	11	1	2	4	10		8	
10	9	12	7	20	5	15	1	2	4	10		8	
11	9	12	8	20	6	17	1	3	6	13		8	
12	9	12	10	20	7	20	1	3	6	13		8	
13	9	12	12	25	8	22	1	4	12	17		8	
14	9	12	14	30	9	25	2	5	15	20		8	
15	9	12	16	35	10	27	3	7	15	20		8	
16	9	12	18	35	12	30	5	8	15	20		8	
17	9	12	18	35	14	30	5	8	15	20		8	
17+	9	12	18	35	14	30	5	8	15	20		8	

Passing = touching fingertips together behind the back

\* Number on left is lower end of HFZ; number on right is upper end of HFZ

\*\* Test scored Pass/Fail; must reach this distance to pass.

©1982, 1986, The Cooper Institute for Aerobics Research, Dallas, Texas.

## GIRLS

Age	One-mile run		PACER		Walk test & $\dot{V}O_2$ max		Percent fat		Body mass index		Curl-up	
	min:sec		# laps		ml/kg/min						# complete	
5							32	17	21	16.2	2	10
6	Completion of distance. Time standards not recommended.		Participation in run. Lap count standards not recommended.				32	17	21	16.2	2	10
7							32	17	22	16.2	4	14
8							32	17	22	16.2	6	20
9							32	17	23	16.2	9	22
10	12:30	9:30	15	41	40	48	32	17	23.5	16.6	12	26
11	12:00	9:00	15	41	39	47	32	17	24	16.9	15	29
12	12:00	9:00	23	41	38	46	32	17	24.5	16.9	18	32
13	11:30	9:00	23	51	37	45	32	17	24.5	17.5	18	32
14	11:00	8:30	23	51	36	44	32	17	25	17.5	18	32
15	10:30	8:00	23	51	35	43	32	17	25	17.5	18	35
16	10:00	8:00	32	61	35	43	32	17	25	17.5	18	35
17	10:00	8:00	41	61	35	43	32	17	26	17.5	18	35
17+	10:00	8:00	41	61	35	43	32	17	27.3	18.0	18	35

Age	Trunk lift		Push-up		Modified pull-up		Pull-up		Flexed arm hang		Back-saver sit & reach**		Shoulder stretch
	inches		# complete		# complete		# complete		seconds		inches		
5	6	12	3	8	2	7	1	2	2	8		9	
6	6	12	3	8	2	7	1	2	2	8		9	
7	6	12	4	10	3	9	1	2	3	8		9	
8	6	12	5	13	4	11	1	2	3	10		9	
9	6	12	6	15	4	11	1	2	4	10		9	
10	9	12	7	15	4	13	1	2	4	10		9	
11	9	12	7	15	4	13	1	2	6	12		10	
12	9	12	7	15	4	13	1	2	7	12		10	
13	9	12	7	15	4	13	1	2	8	12		10	
14	9	12	7	15	4	13	1	2	8	12		10	
15	9	12	7	15	4	13	1	2	8	12		12	
16	9	12	7	15	4	13	1	2	8	12		12	
17	9	12	7	15	4	13	1	2	8	12		12	
17+	9	12	7	15	4	13	1	2	8	12		12	

Passing = touching fingertips together behind the back

\* Number on left is lower end of HFZ; number on right is upper end of HFZ

\*\*Test scored Pass/Fail; must reach this distance to pass.

©1992, 1999, The Cooper Institute for Aerobics Research, Dallas, Texas.

## Appendix G

### Contingency Tables by Age and Gender.

Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Eight-Year-Old Boys (n=46).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	5	9
	Unhealthy	1	31

		FAH	
		Healthy	Unhealthy
PSU	Healthy	11	3
	Unhealthy	6	26

		PU	
		Healthy	Unhealthy
PSU	Healthy	12	2
	Unhealthy	16	16

Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Nine-Year-Old Boys (n=50).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	5	11
	Unhealthy	0	34

		FAH	
		Healthy	Unhealthy
PSU	Healthy	8	8
	Unhealthy	4	30

		PU	
		Healthy	Unhealthy
PSU	Healthy	14	2
	Unhealthy	17	17



Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Ten-Year-Old Boys (n=61).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	14	15
	Unhealthy	1	31

		FAH	
		Healthy	Unhealthy
PSU	Healthy	18	11
	Unhealthy	6	26

		PU	
		Healthy	Unhealthy
PSU	Healthy	26	3
	Unhealthy	13	19

Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Eleven-Year-Old Boys (n=44).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	12	4
	Unhealthy	2	26

		FAH	
		Healthy	Unhealthy
PSU	Healthy	16	0
	Unhealthy	11	17

		PU	
		Healthy	Unhealthy
PSU	Healthy	16	0
	Unhealthy	13	15

Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Eight-Year-Old Girls (n=39).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	3	18
	Unhealthy	0	18

		FAH	
		Healthy	Unhealthy
PSU	Healthy	11	10
	Unhealthy	7	11

		PU	
		Healthy	Unhealthy
PSU	Healthy	19	2
	Unhealthy	9	9

Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Nine-Year-Old Girls (n=56).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	5	28
	Unhealthy	1	22

		FAH	
		Healthy	Unhealthy
PSU	Healthy	18	15
	Unhealthy	5	18

		PU	
		Healthy	Unhealthy
PSU	Healthy	32	1
	Unhealthy	12	11

Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Ten-Year-Old Girls (n=44).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	11	18
	Unhealthy	0	15

		FAH	
		Healthy	Unhealthy
PSU	Healthy	21	8
	Unhealthy	3	12

		PU	
		Healthy	Unhealthy
PSU	Healthy	28	1
	Unhealthy	7	8

Agreement Between the Push-up Test and Modified Pull-Up, Flexed-Arm Hang, and Pull-Up Tests for Eleven-Year-Old Girls (n=43).

		MPU	
		Healthy	Unhealthy
PSU	Healthy	5	17
	Unhealthy	1	20

		FAH	
		Healthy	Unhealthy
PSU	Healthy	20	2
	Unhealthy	12	9

		PU	
		Healthy	Unhealthy
PSU	Healthy	21	1
	Unhealthy	13	8

**Appendix H**

**Survival Curve Performances for Eight and Nine, Nine and Ten, and Ten  
and Eleven-Year-Old Boys and Girls on the Tests of Strength and  
Endurance.**

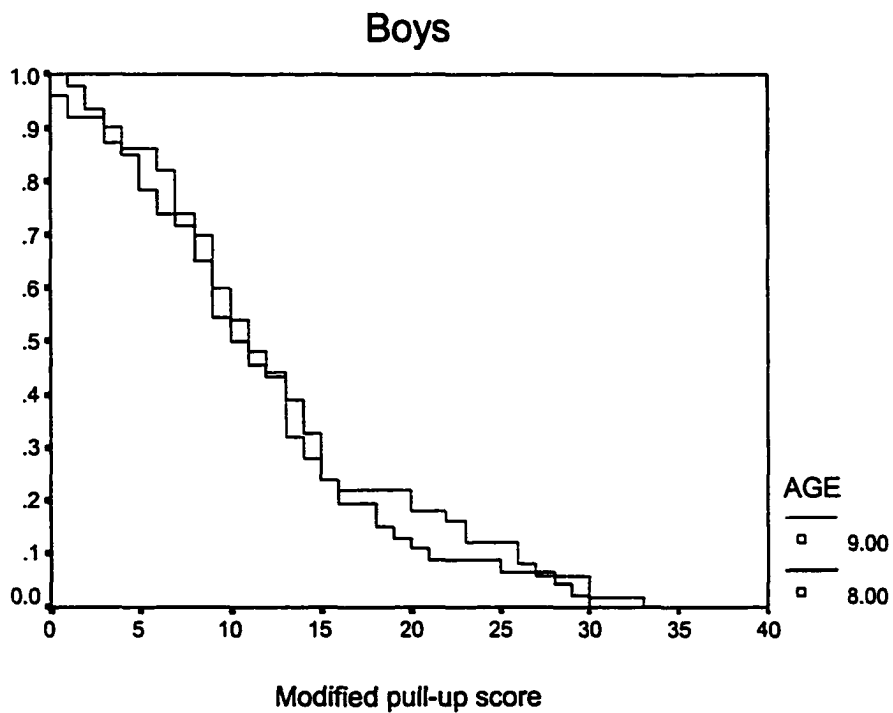
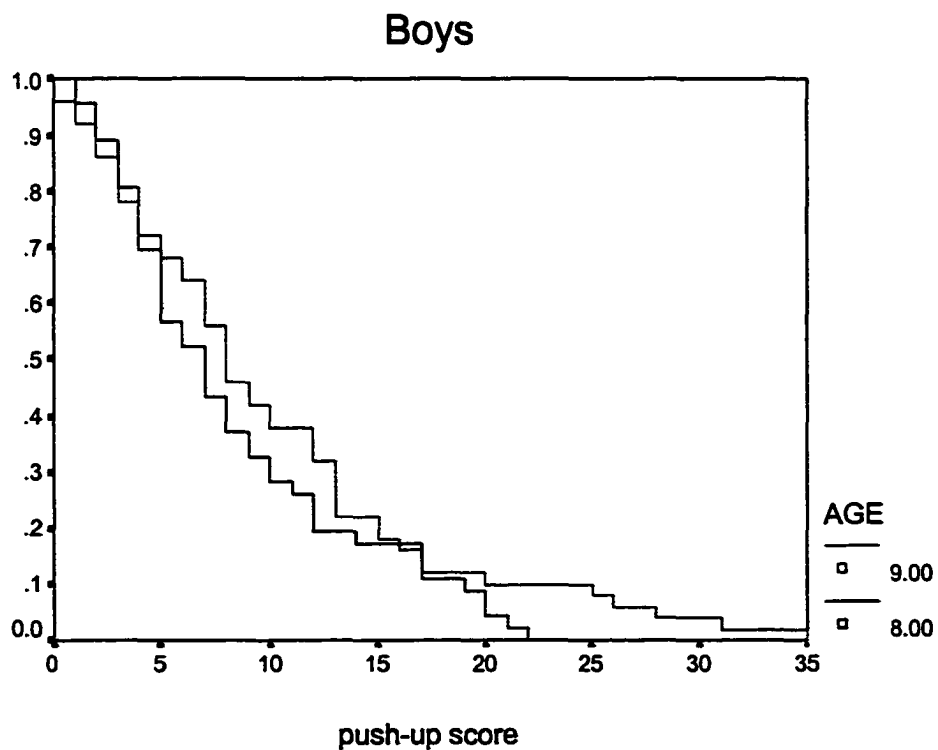
Figure 3Figure 4



Figure 5

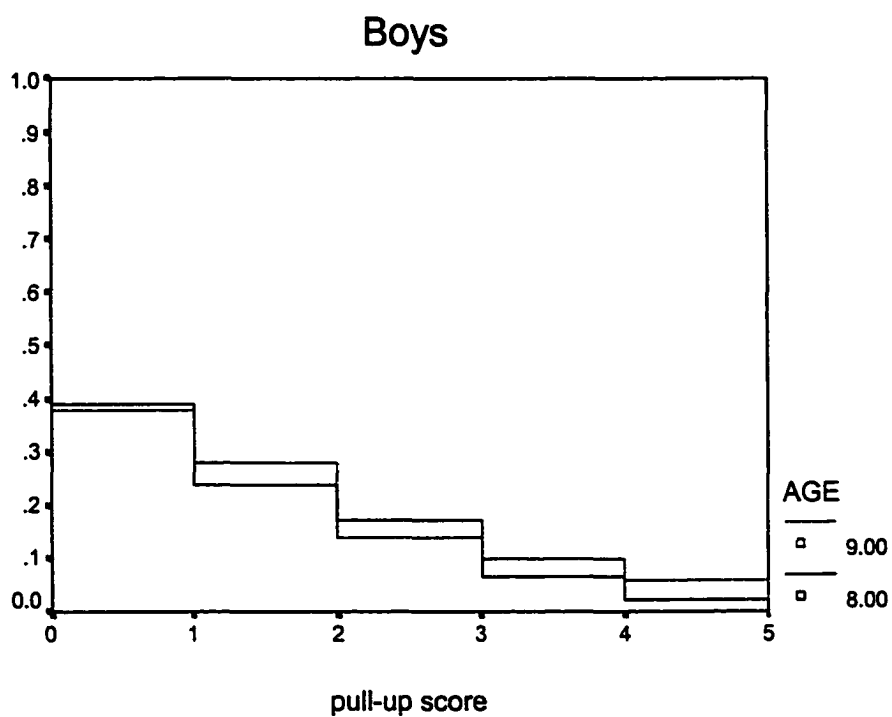
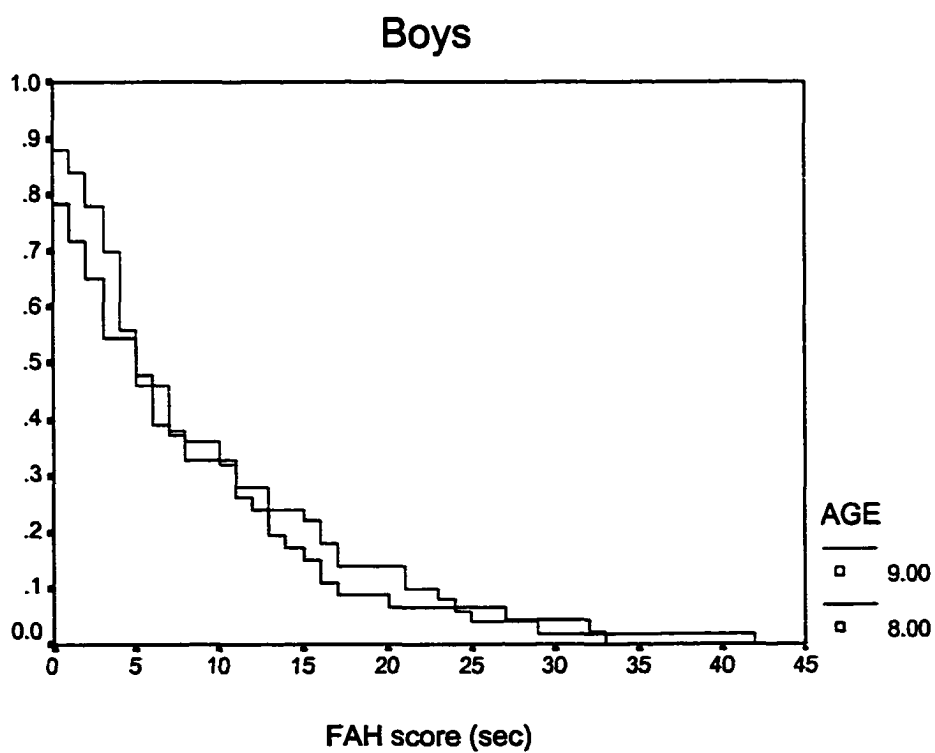


Figure 6



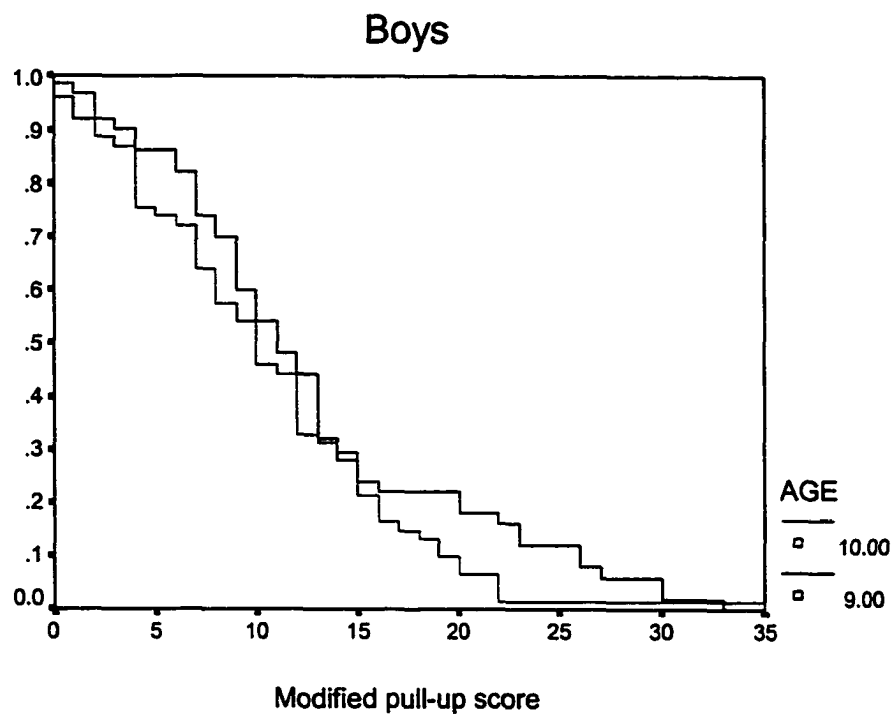
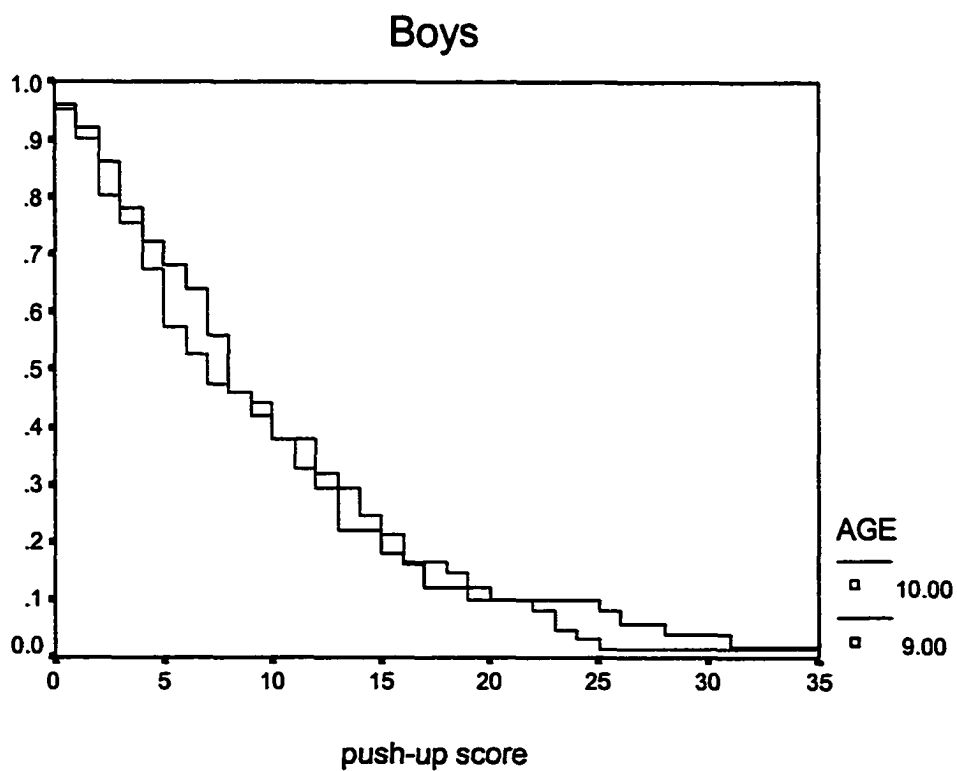
**Figure 7****Figure 8**

Figure 9

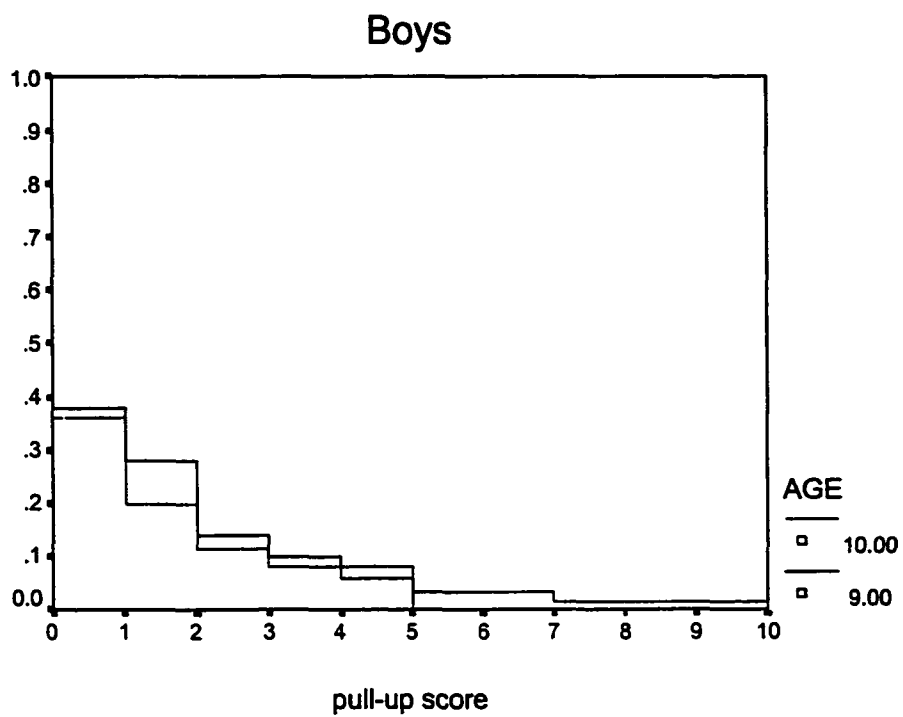
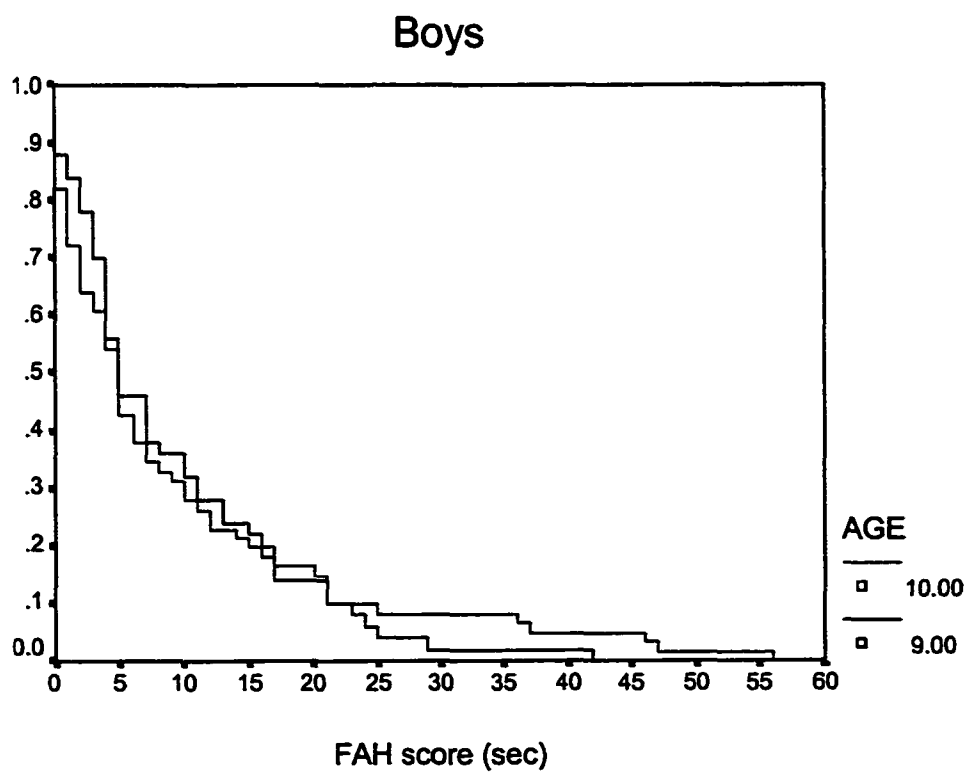
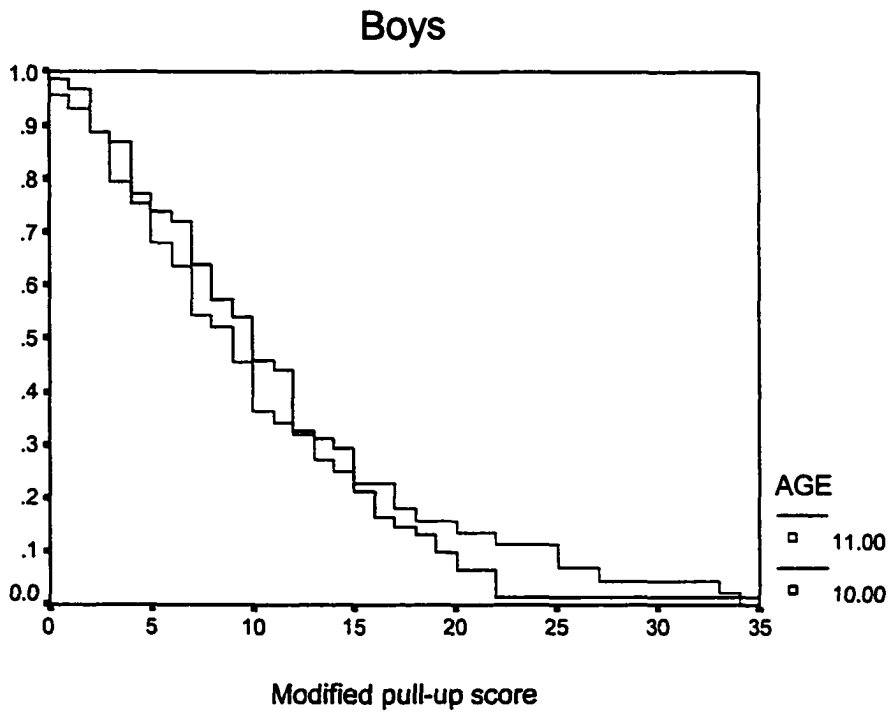


Figure 10



**Figure 11**



**Figure 12**

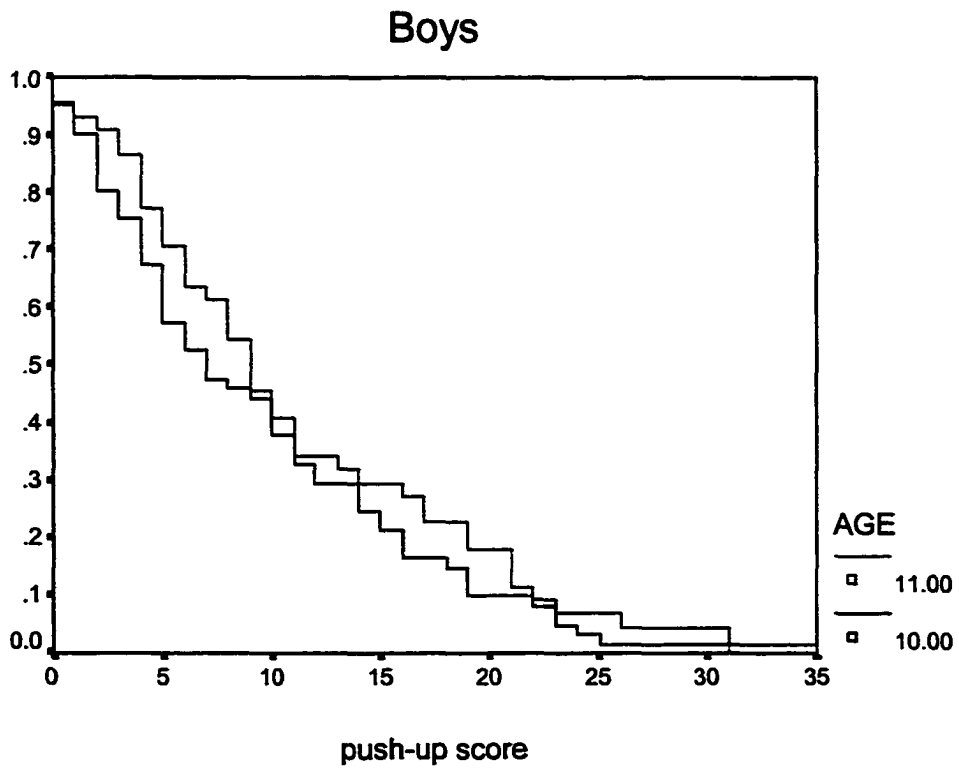


Figure 13

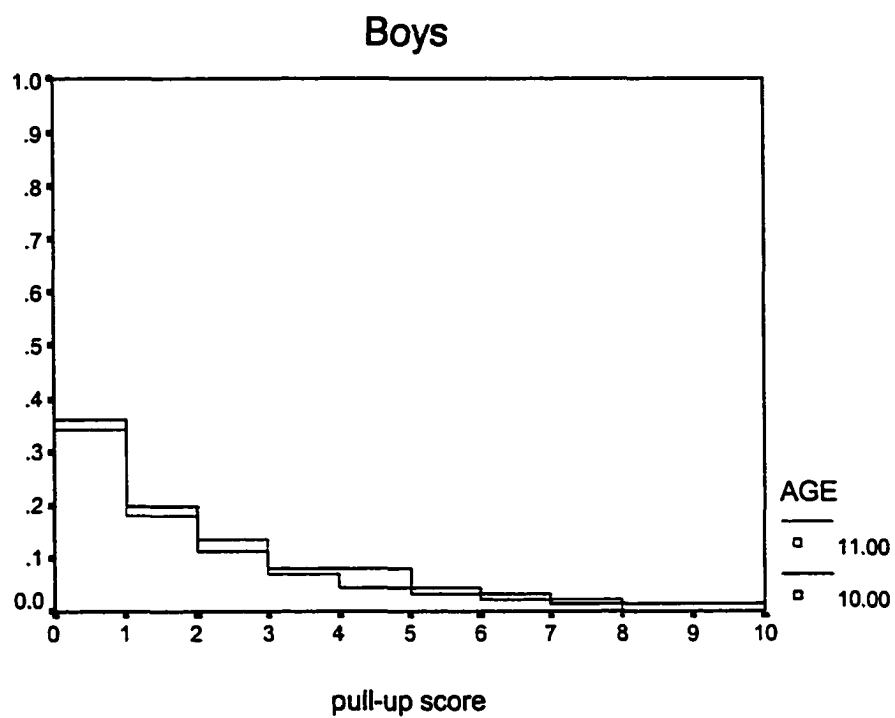


Figure 14

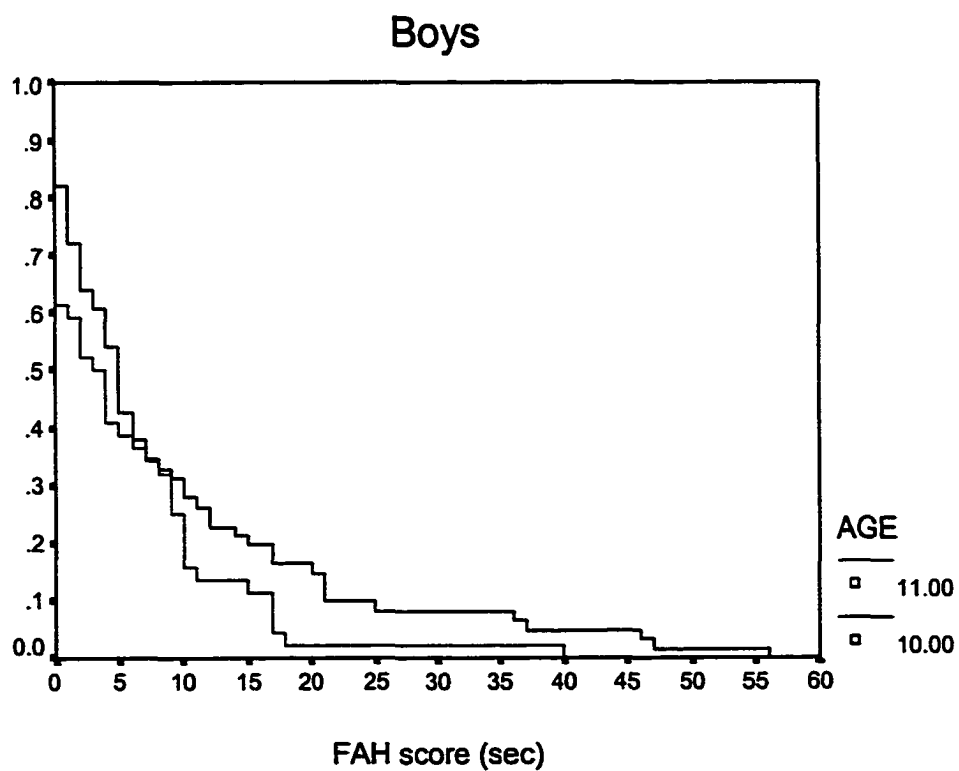


Figure 15

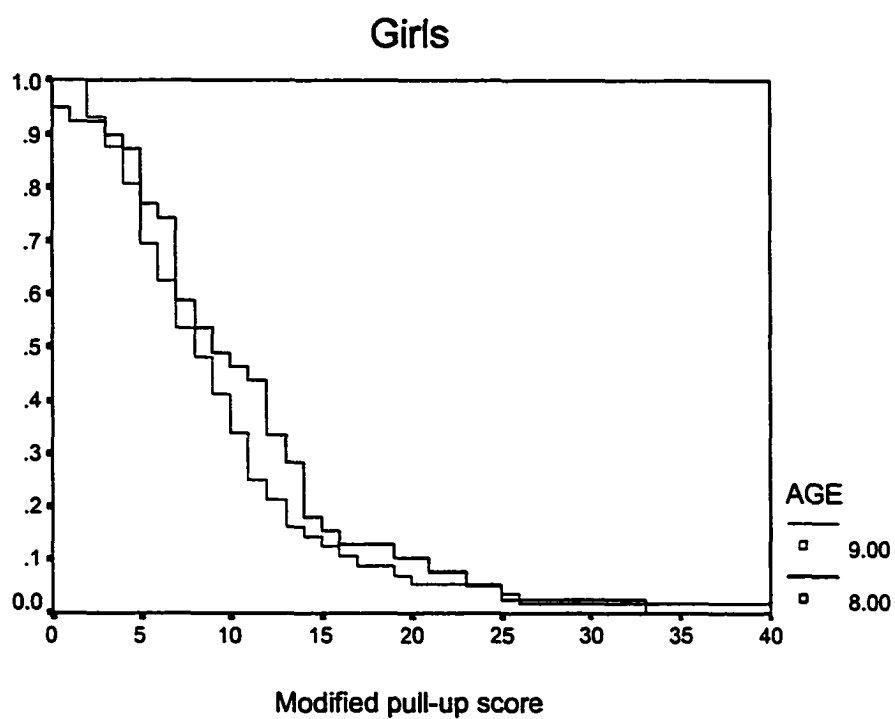


Figure 16

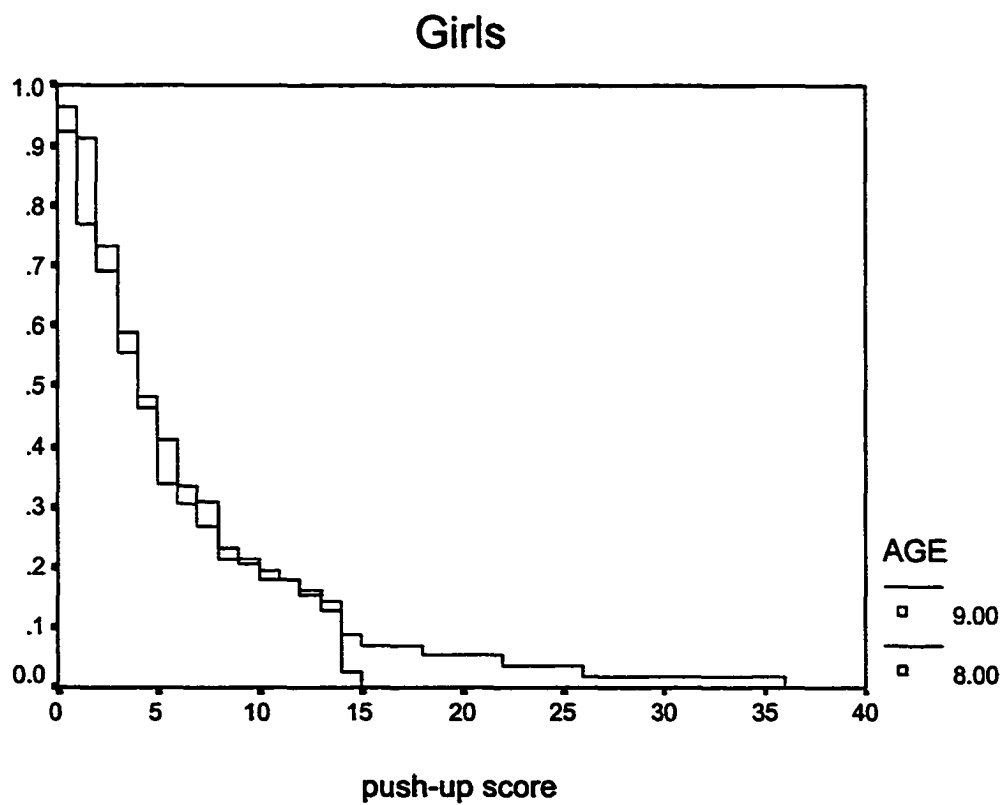


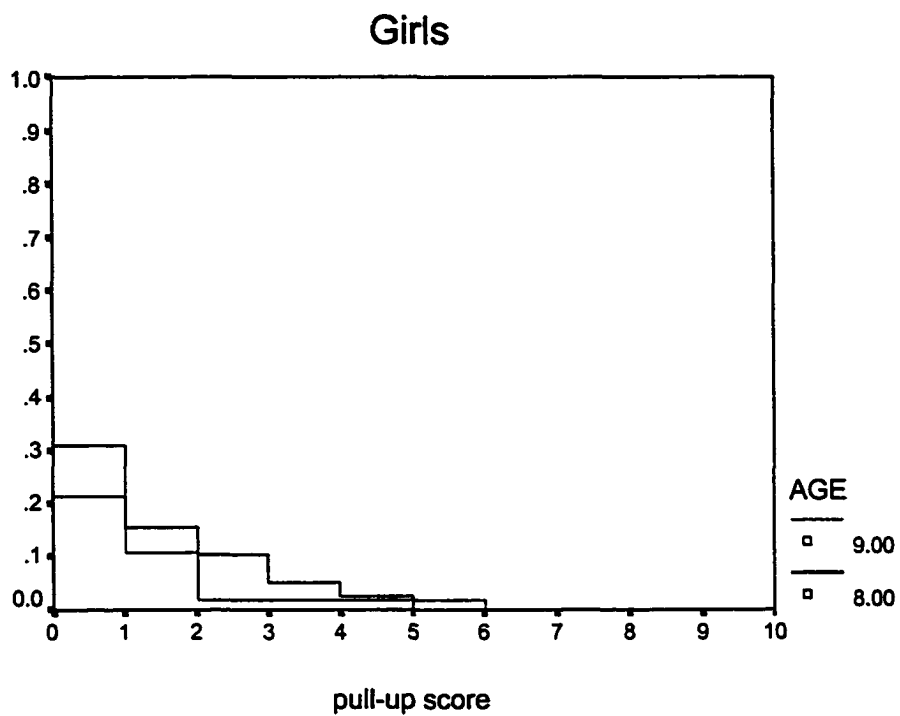
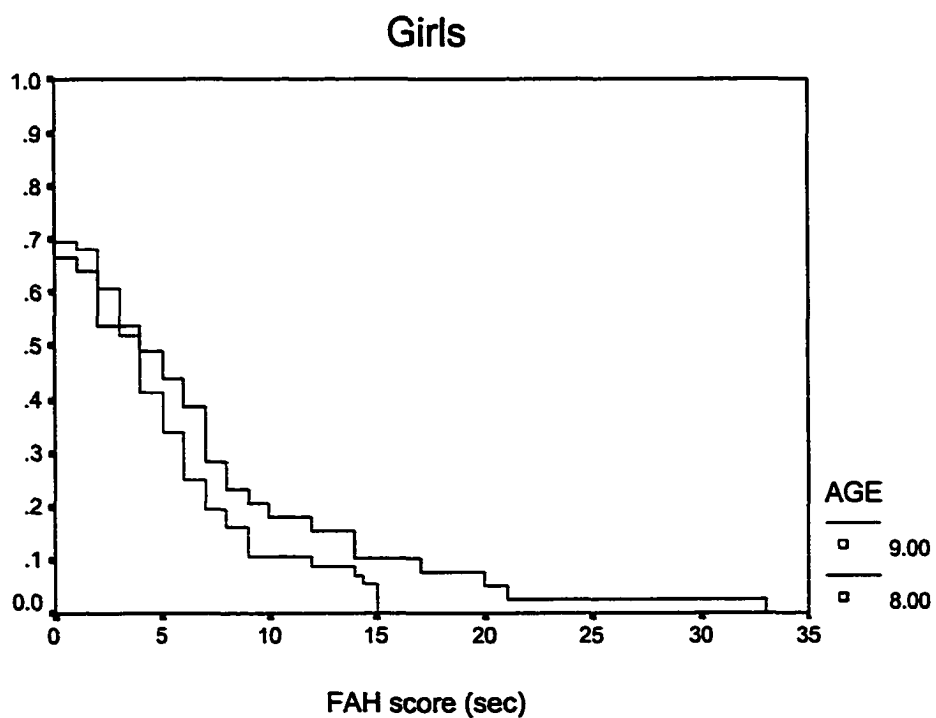
Figure 17Figure 18

Figure 19

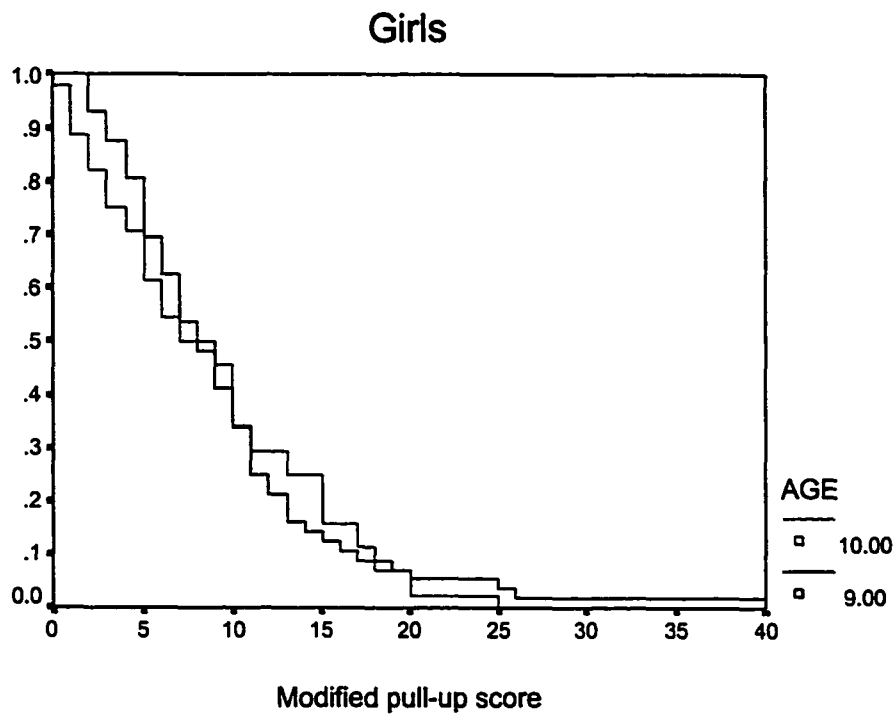


Figure 20

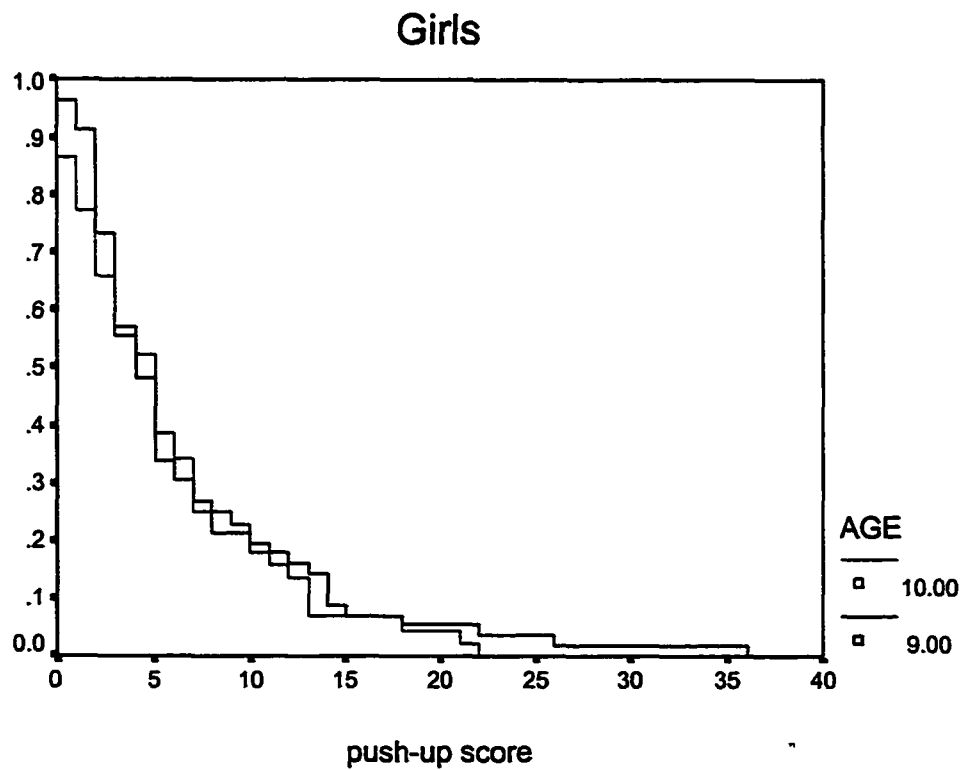




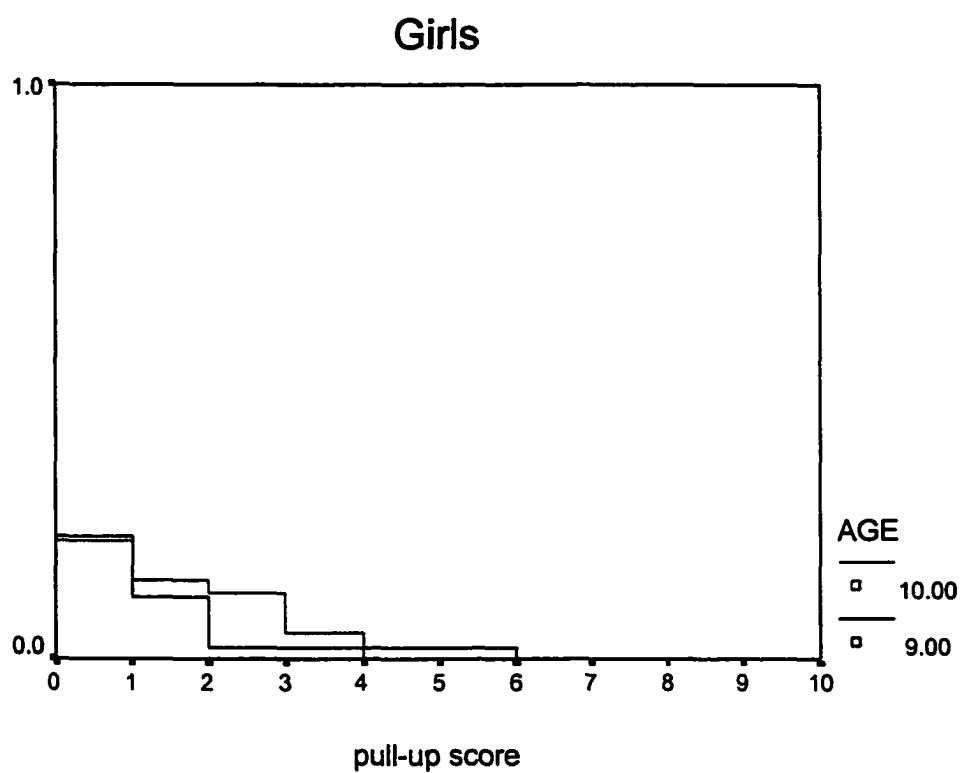
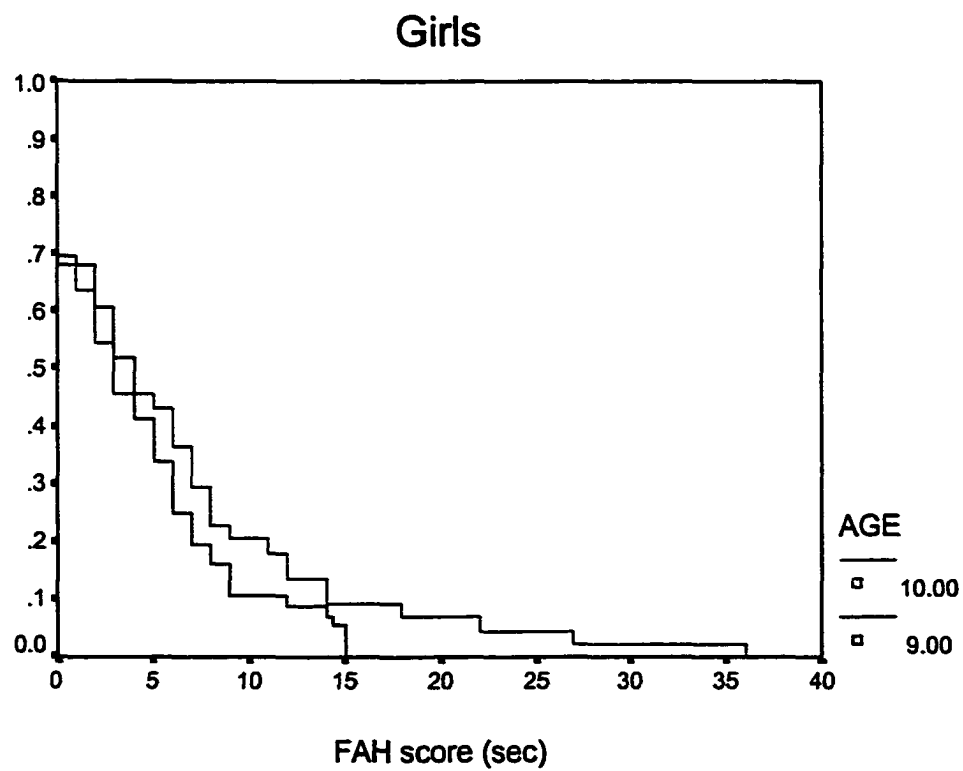
Figure 21Figure22

Figure 23

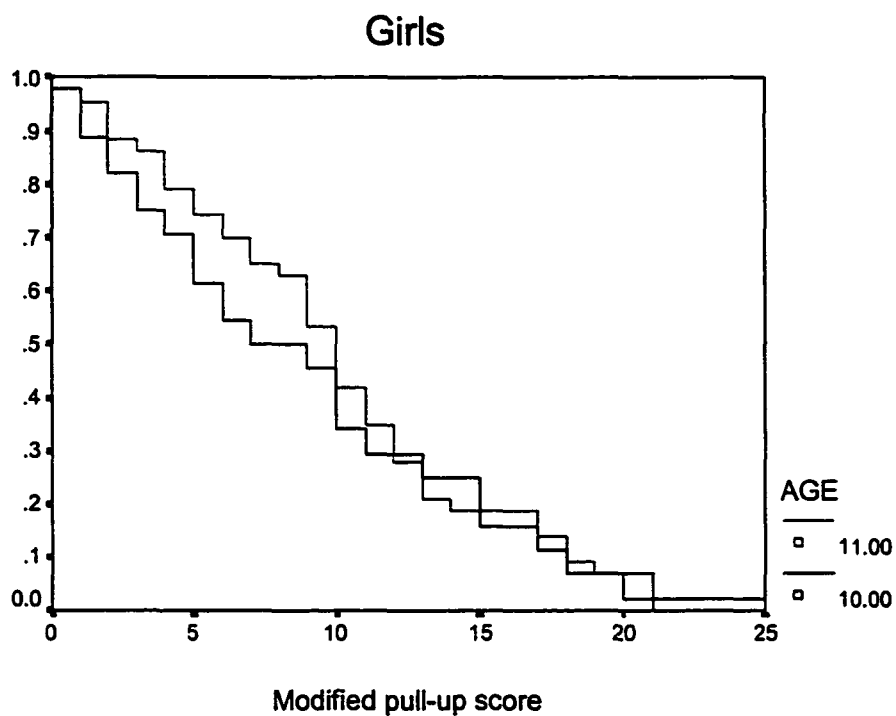


Figure 24

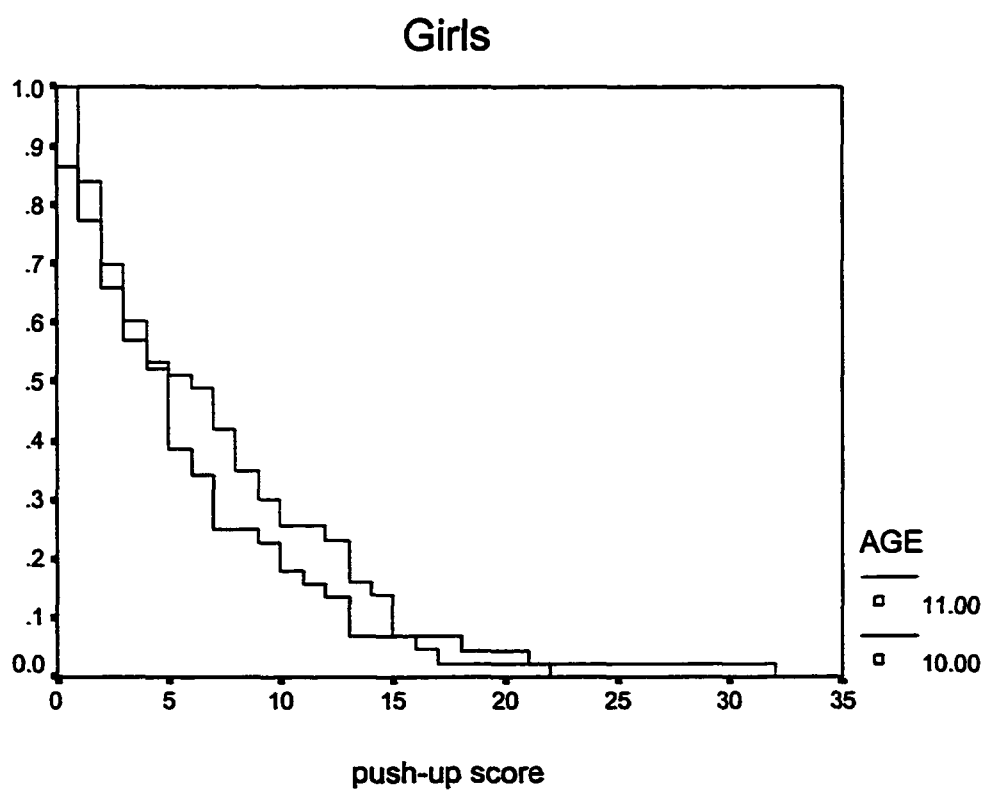
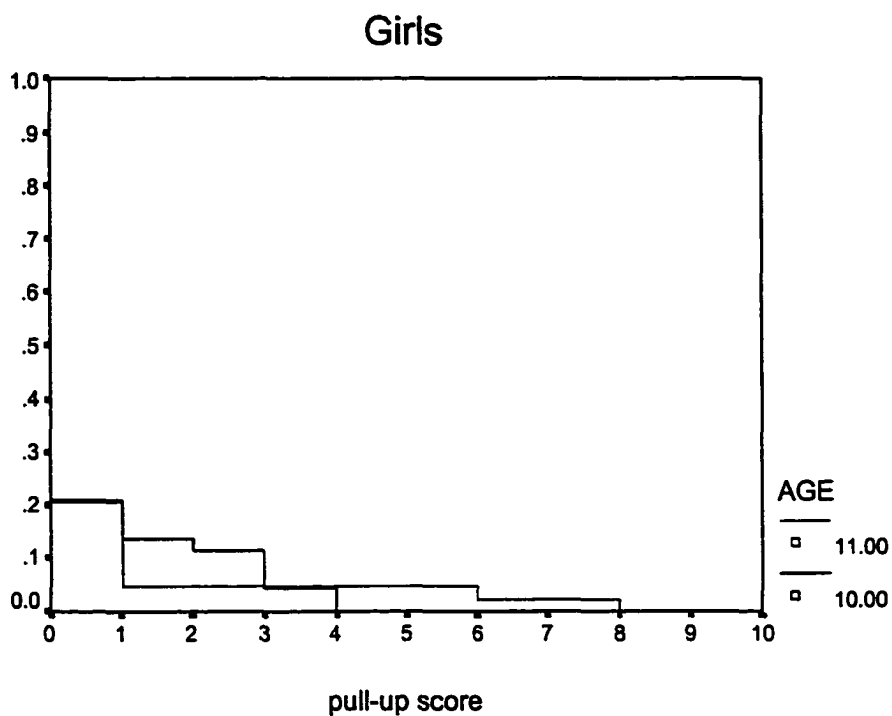
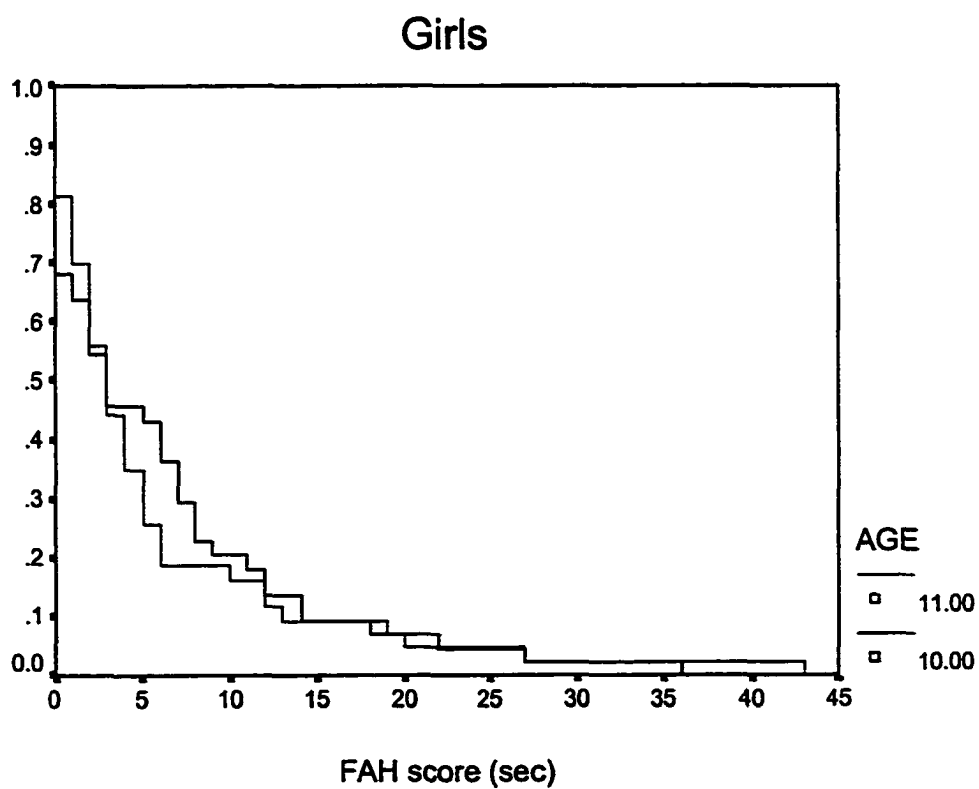


Figure 25Figure 26

Appendix I

Survival Curve Performances Between Eight, Nine, Ten, and  
Eleven-Year-Old Boys and Girls Tests of Strength and Endurance.

Figure 27

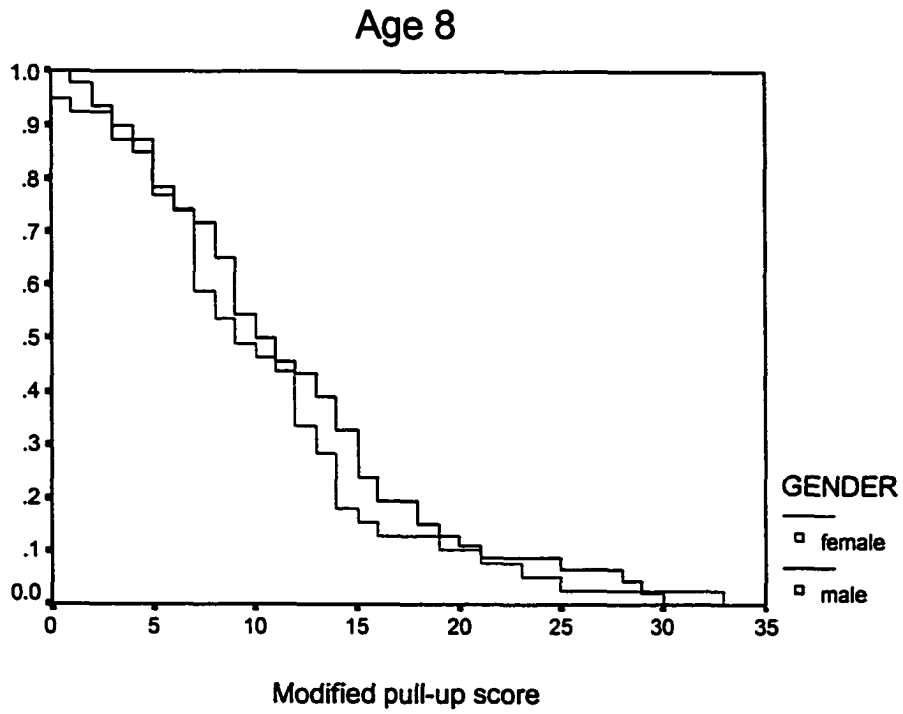


Figure 28

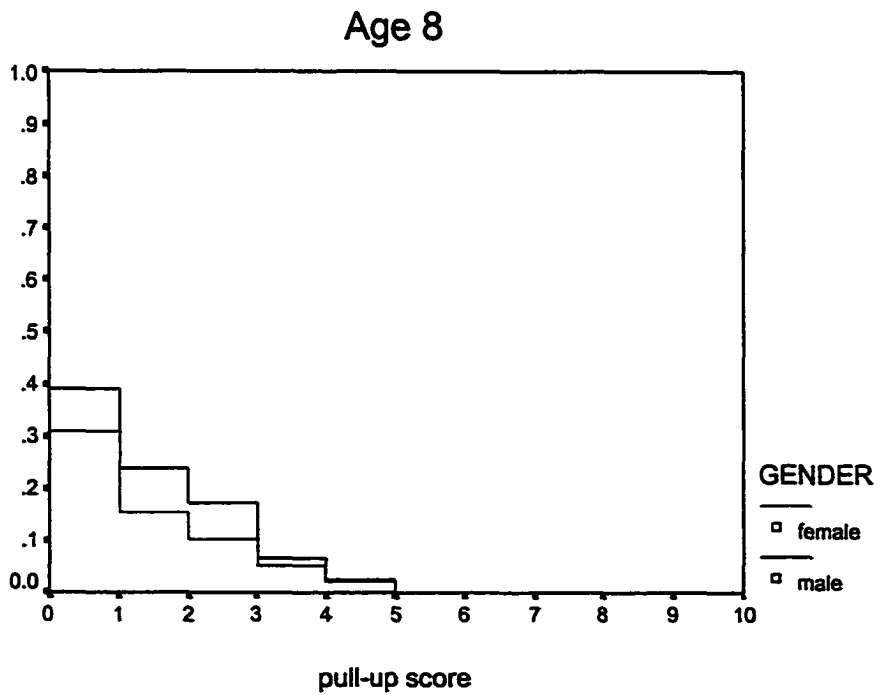


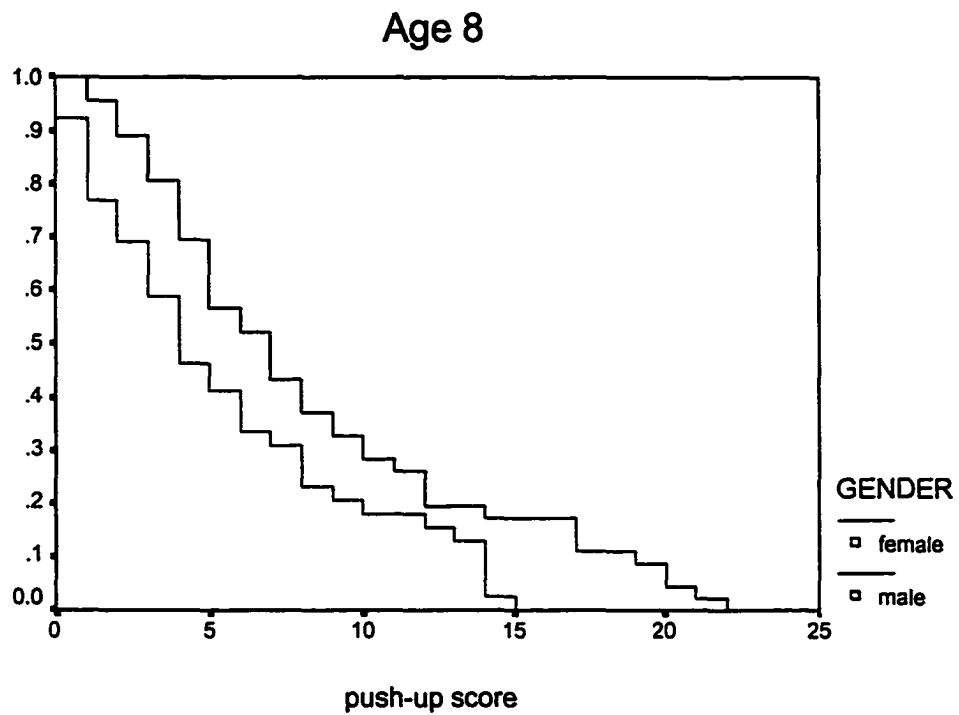
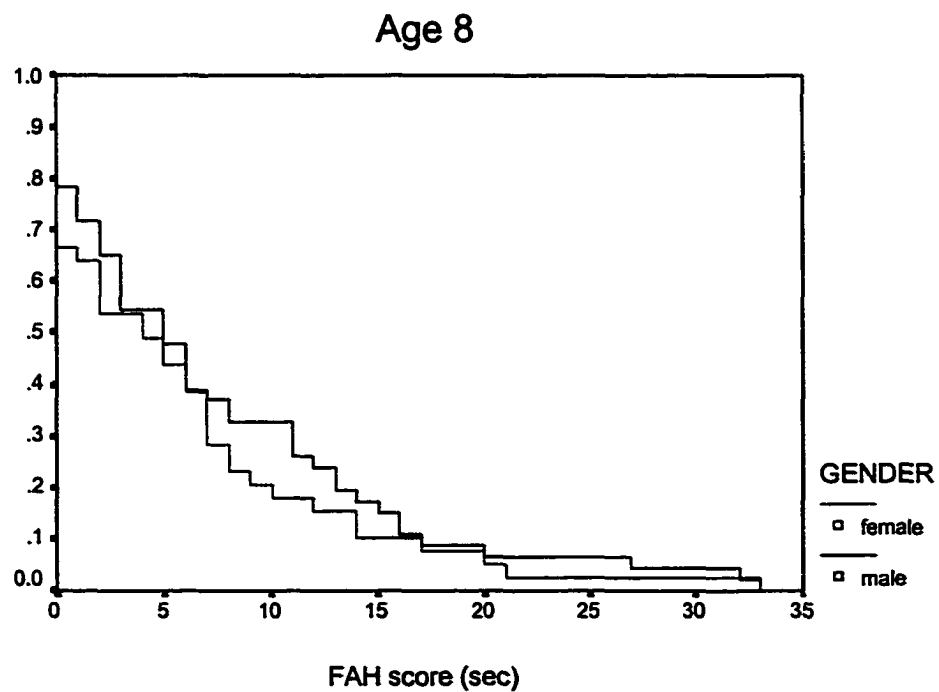
Figure 29Figure 30

Figure 31

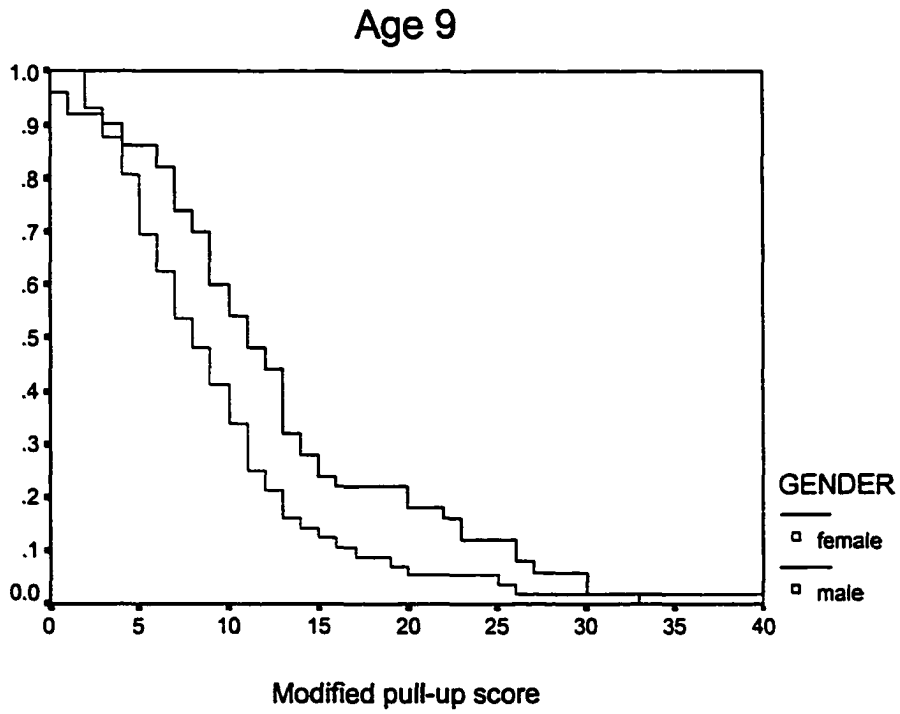
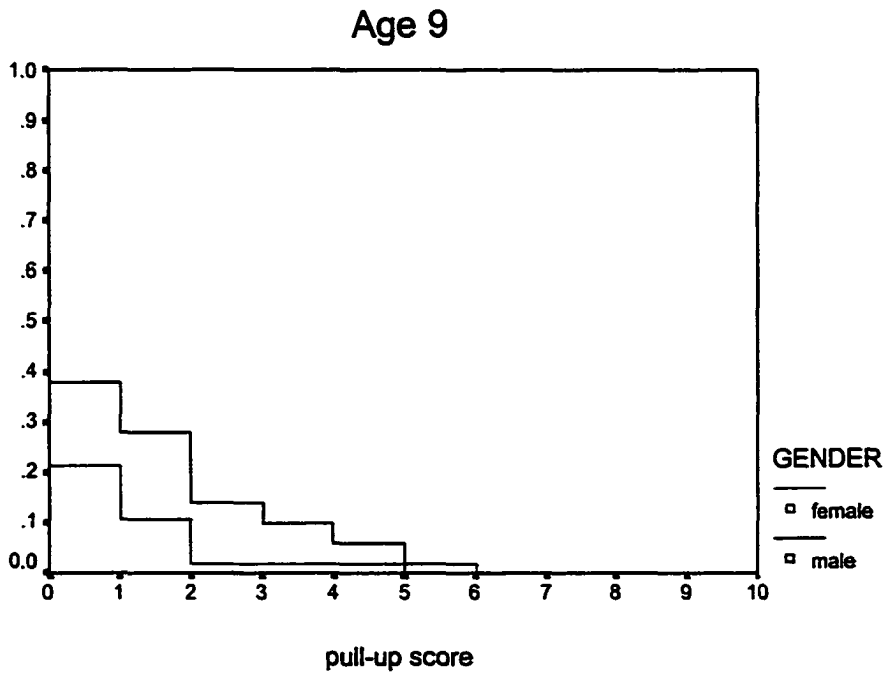
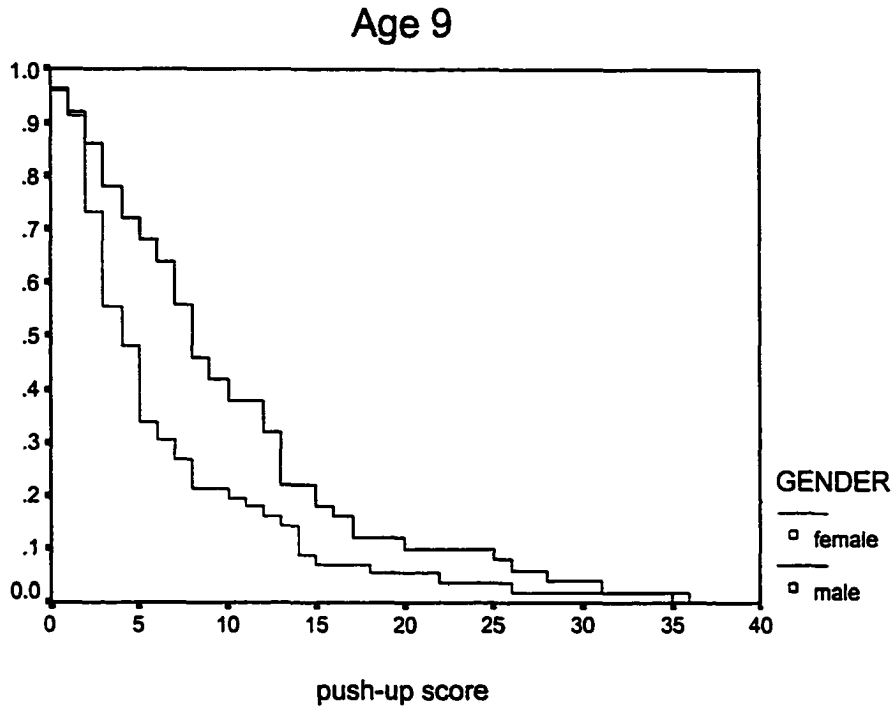


Figure 32



**Figure 33**



**Figure 34**

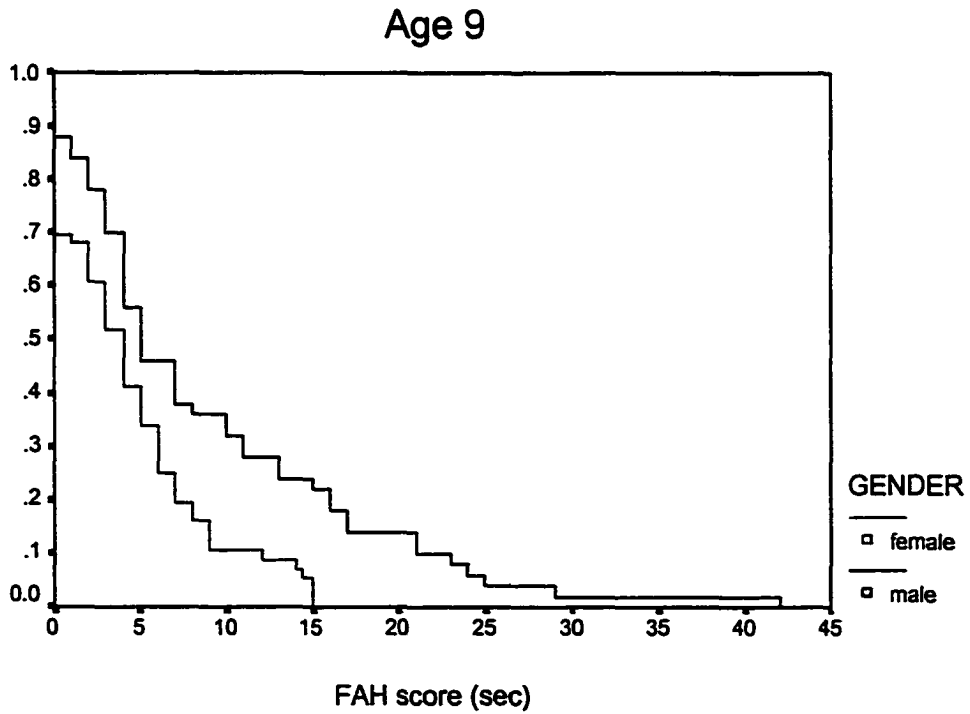




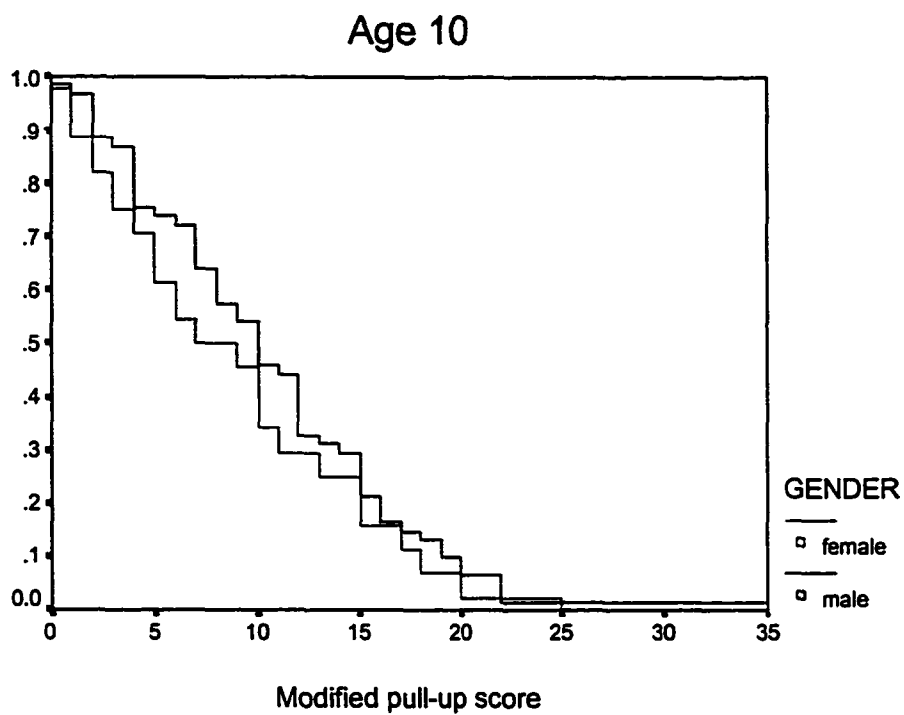
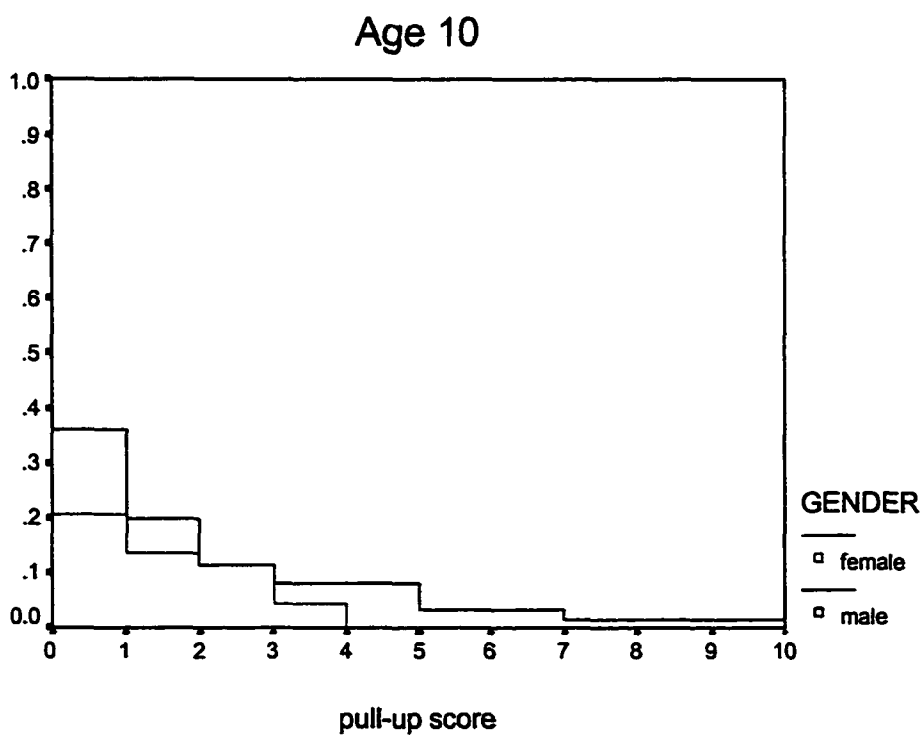
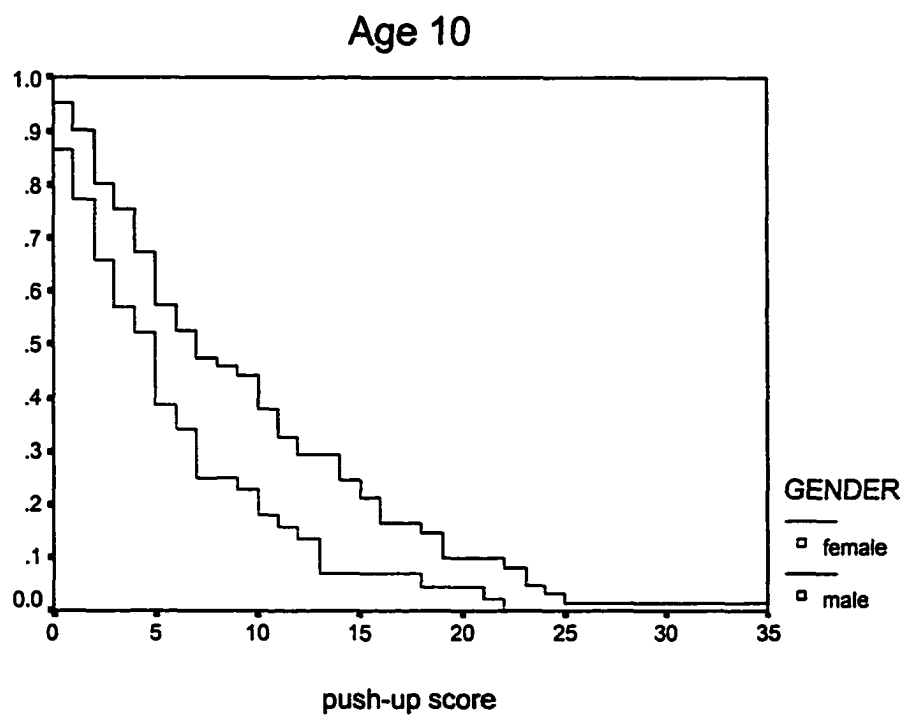
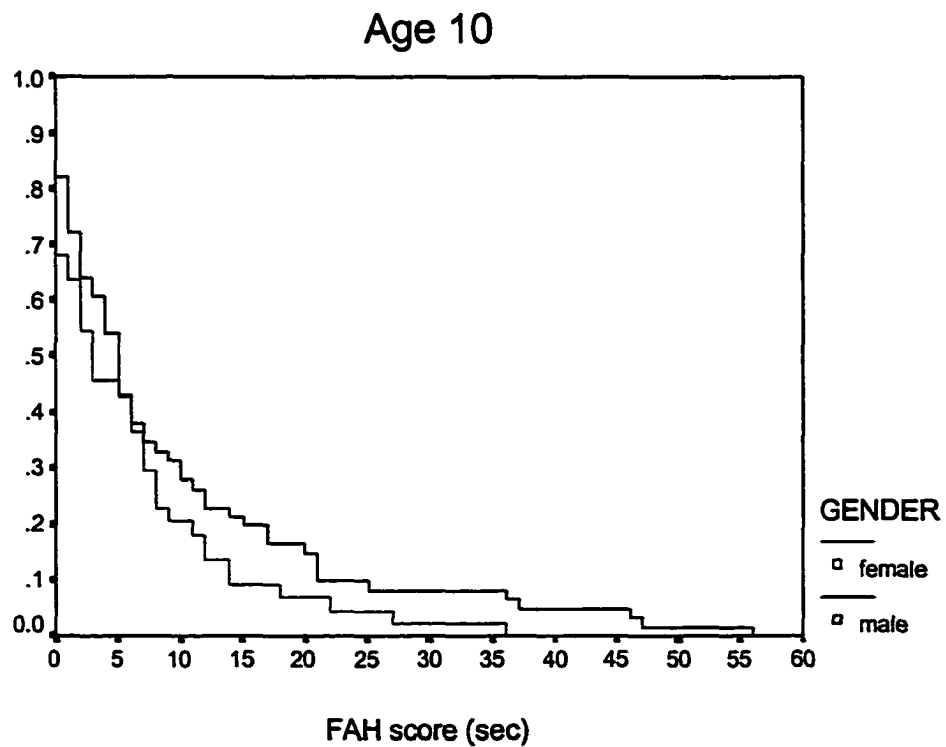
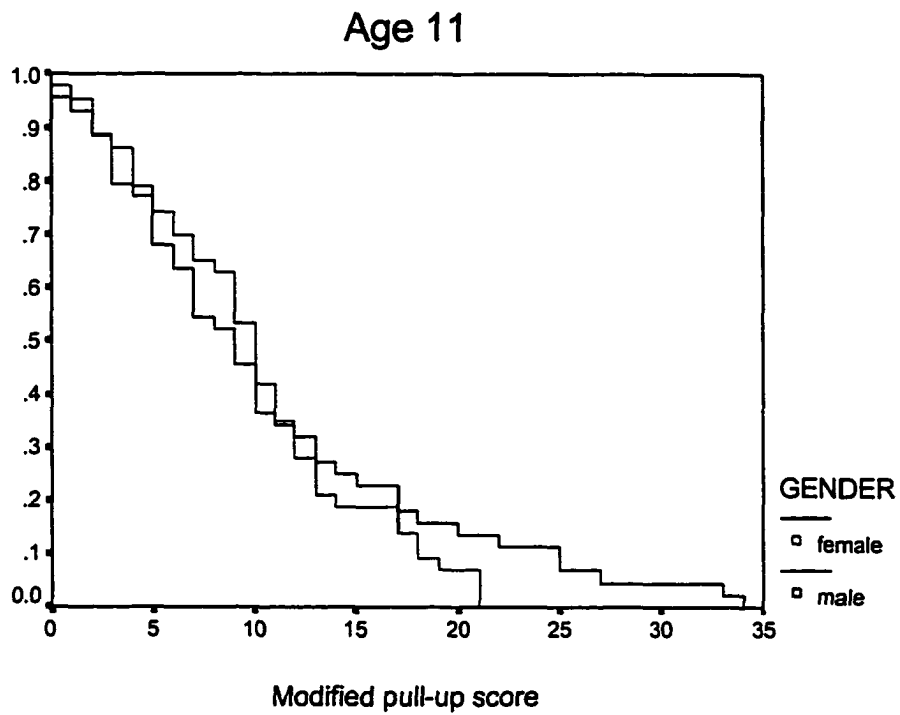
Figure 35Figure 36

Figure 37Figure 38

**Figure 39**



**Figure 40**

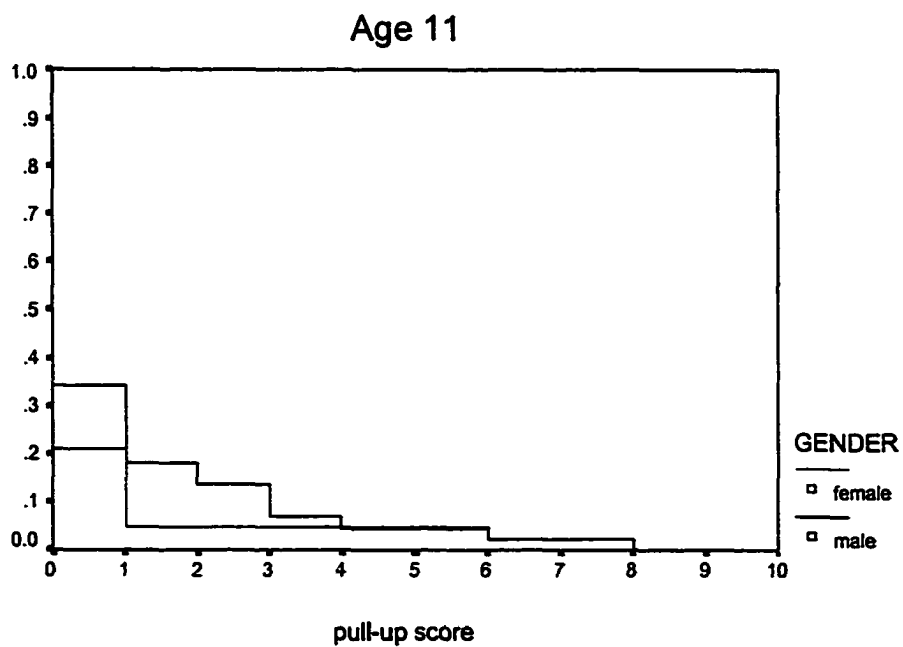
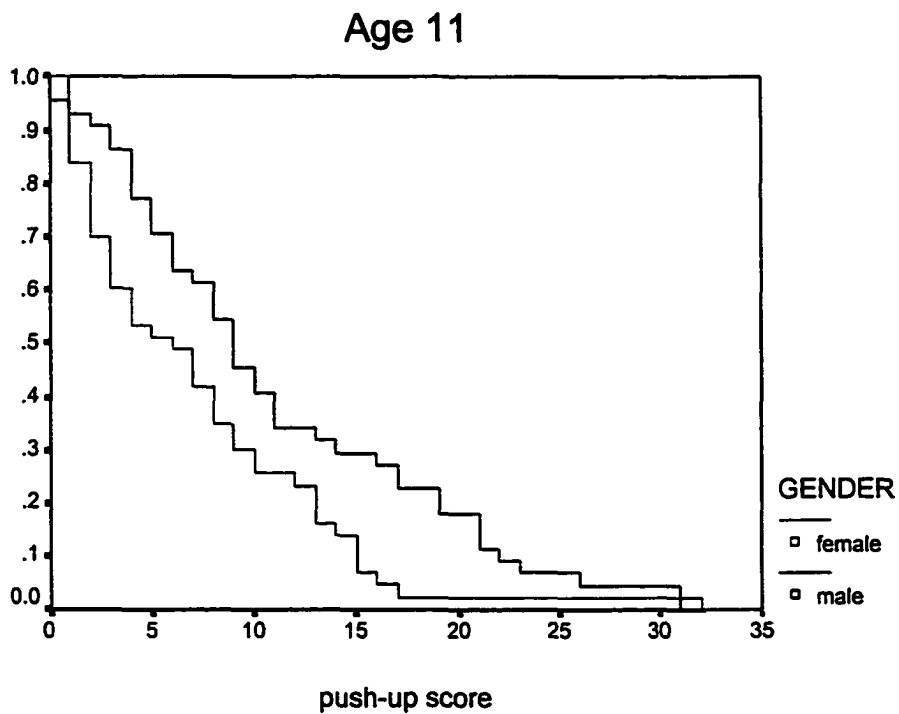
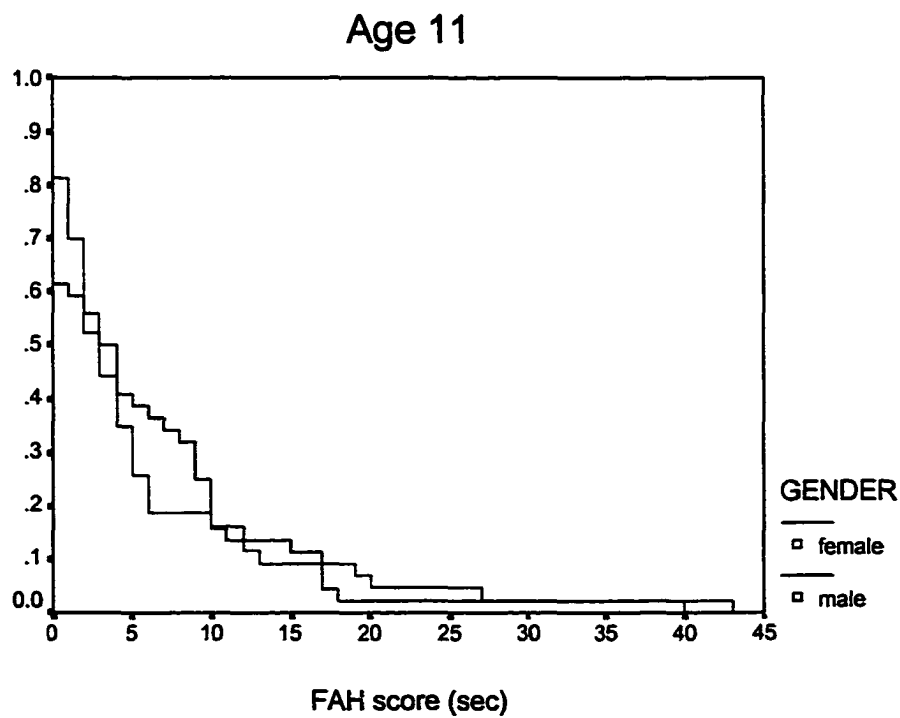


Figure 41Figure 42

## REFERENCES

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

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

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

Chuang, J. H. (2001). Agreement between categorical measurements: Kappa statistics [On-line]. Available:  
<http://www.cpmc.columbia.edu/homepages/chuangj/kappa/>

Chun, D. M., Corbin, C. B., & Pangrazi, R. P. (2000). Validation of criterion-referenced standards for the mile run and progressive aerobic cardiovascular endurance tests. Research Quarterly for Exercise and Sport, 71(2), 125-134.

Cooper Institute for Aerobics Research. (1999). The Prudential FITNESSGRAM Test Administration Manual. Dallas, TX: Author.

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

Cureton, K. J. & Warren, G. L. (1990). Criterion-referenced standards for youth health-related tests: A tutorial. Research Quarterly for Exercise and Sport, 61(1), 7-19.

Engelman, M. E., & Morrow, J. R., Jr. (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.

Greenhouse, J. B., Stangl, D., & Bromberg, J. (1989). An introduction to survival analysis: Statistical methods for analysis of clinical trial. Journal of Consulting and Clinical Psychology, 57(4), 536-544.

Gruber, F. A. (1999). Tutorial: Survival analysis- A statistic for clinical, efficacy, and theoretical applications. Journal of Speech, Language, and Hearing Research, 42, 432-447.

Hamill, P. V. V., Drizd, T. A., Johnson, C. L., Reed, R. B., Roche, A. F., & Moore, W. M., (1979). Physical growth: National center for health statistics percentiles. American Journal of Clinical Nutrition, 32, 607-629.

Jackson, A., Bruya, L., Baun, W., Richardson, P., Weinberg, R., and Caton, I. (1992). Baumgartner's modified pull-up test for male and female elementary school aged children. Research Quarterly for Exercise and Sport, 53(2), 163-164.

Jackson, A. W., Fromme, C., Plitt, H., & Mercer, J. (1994). Reliability and validity of a 1-minute push-up test for young adults. Research Quarterly for Exercise in Sport, 65(Suppl.), A57-A58. (Abstract).

Kleinbaum, K. G. (1995). Survival analysis: A self-learning text. New York: Springer.

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

Looney, M. A., & Plowman, S. A. (1990). Passing rates of American children and youth on the FITNESSGRAM criterion-referenced physical fitness standards. Research Quarterly for Exercise and Sport, 61, 215-223.

Mahar, M. T., Rowe, D. A., Parker, C. R., Mahar, F. J., Dawson, D. M., & Holt, J. E. (1997). Criterion-referenced and norm-referenced agreement between the mile run/walk and PACER. Measurement in Physical Education and Exercise Science, 1,(4), 245-258.

Malina, R. M. & Bouchard, C. (1991). Growth, Maturation, and Physical Activity. Human Kinetics: Champaign, IL.

McManis, B. G., & Wuest, D. A. (1994). Stability reliability of the modified push-up in children. Research Quarterly for Exercise in Sport, 65 (Suppl.), A58-A59. (Abstract).

Morrow, J. R., Jr., Jackson, A. W., Disch, J. G., & Mood, D. P. (1995). Measurement and evaluation in human performance. Champaign, IL: Human Kinetics.

Pate, R. R., Burgess, M. L., Woods, J. A., Ross, J. G., & Baumgartner, T. (1993). Validity of field tests of upper body muscular strength. Research Quarterly for Exercise and Sport, 64, 17-24.

Pate, R.R.& Hohn, R.C. (1994). Health and fitness through physical education (pp. 215-217). Champaign, IL: Human Kinetics.

Pate, R. R., Ross, J.G., Baumgartner, T. A., & Sparks, E. (1987). The modified pull-up test. Journal of Physical Education, Recreation & Dance, 58(9), 71-73.

Pate, R. R. & Shepard, R. J. (1989). Characteristics of physical fitness in youth. In C. Gisolfi, & D. Lamb (Eds.), Perspectives in exercise science and sport medicine: Vol 2. Youth, exercise, & sport (pp. 1-45). Indianapolis, IN: Benchmark Press.

Rosato, F. D. (1990). Fitness and wellness: The physical connection (2<sup>nd</sup> ed.). New York. West Publishing Company.

Ross, J. G. & Gilbert, G. G. (1985). The national children and youth fitness study: A summary of findings. Journal of Physical Education, Recreation, and Dance, 56, (1), 45-50.

Ross, J. G. & Pate, R. R. (1987). The national children and youth fitness study II: A summary of findings. Journal of Physical Education, Recreation, and Dance, 58(9), 51-56.

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

Rutherford, W. J., & Corbin, C. B. (1994 ). Validation of criterion-referenced standards for tests of arm and shoulder girdle strength and endurance. Research Quarterly for Exercise and Sport, 65, 110-119.

Safrit, M. J. (1990). Introduction to measurement in physical education and exercise science. (2<sup>nd</sup> ed.). Boston. Times Mirror/Mosby College Publishing.

Safrit, M. J. & Wood, T. M. (1995). Validity and reliability of criterion-referenced tests. In M. Safrit & T Woods (Eds.), Introduction to measurement in physical education and exercise science (pp. 174-191). New York: McGraw-Hill.

Walker, J. L., Lloyd, L. K., Bishop, P. A., & Richardson, M. T. (2000). The influence of body size and composition on the successful completion of the FITNESSGRAM pull-up test in fifth-and sixth-grade children. Research Quarterly for Exercise and Sport, 71 (Suppl.), A-54. (Abstract).



Woods, J. A., Burgess, M. L., & Pate, R. R. (2000). Validity of the field tests of upperbody strength in children. Medicine and Science in Sports and Exercise, (suppl), S112.