DEVELOPMENT AND INITIAL VALIDATION OF THE FURTADO-GALLAGHER COMPUTERIZED OBSERVATIONAL MOVEMENT PATTERN ASSESSMENT SYSTEM – FG-COMPASS

by

Ovande Furtado, Jr.

BS, Federal University of Parana, 1997

MS, University of Pittsburgh, 2004

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Ovande Furtado, Jr.

It was defended on

July 29, 2009

and approved by

Louis A Pingel, PhD, Associate Professor, Psychology in Education

Mary E Duquin, PhD, Faculty Emeritus, Health and Physical Activity

Sara R Piva, PhD, Associate Professor, Physical Therapy

Dissertation Advisor: Jere D Gallagher, PhD, Associate Professor, Health and Physical Activity

DEVELOPMENT AND INITIAL VALIDATION OF THE FURTADO-GALLAGHER COMPUTERIZED OBSERVATIONAL MOVEMENT PATTERN ASSESSMENT SYSTEM – FG-COMPASS

Ovande Furtado, Jr., PhD

University of Pittsburgh, 2009

Mastery in fundamental movement skill (FMS) performance (e.g., kicking, jumping, throwing) has been considered an important factor in preventing unhealthy weight gain (Okely, Booth & Chey, 2004); as well as helping increases in participation of organized and habitual physical activity (Foley, Harvey, Chun & Kim, 2008; Hume, Okely, Bagley, Telford, Booth, Crawford & Salmon, 2008; Mazzardo, 2008; Okely, Booth & Chey, 2004; Okely, Booth & Patterson, 2001) among children and adolescents. Thus, assessing FMS development becomes crucial in school settings. Therefore, the purpose of this study was to develop and collect initial validity evidence for a new observational assessment tool (FG-COMPASS) to evaluate FMS development of children 5- to 10 years of age. In Phase I of this study, 110 video clips of children performing 5 locomotor and 6 object-control FMS were developed. In Phase II, the rating scales (composite decision trees) were developed for each FMS. In addition, the efficacy of the decision trees was evaluated by comparing judgments of 30 undergraduate students with a standard. Weighted kappa indicated that the agreement was best for hop (K_w= .85), followed by strike and batting (K_w= .79), skip (K_w= .77), overhand throw (K_w= .74), catch and hand dribble (K_w= .72), and horizontal jump (K_w= .70). The poorest agreement occurred in the skills of kick (K_w= .51), and side slide and leap (K_w= .61). The proportion of specific agreement (P_s) was calculated for each skill with the purpose to find out the source of disagreement. Skills that had at least one category

(e.g., initial, elementary, mature) with P_s values below .70 were further inspected. Six skills were selected for further analysis (side slide, horizontal jump, leap, kick, hand dribble, and overhand throw). The decision trees for all six skills underwent modifications. In conclusion, this study provided initial validity evidence that the decision trees (rating scale) developed for the FG-COMPASS could be used to classify individuals based on their FMS development. However, reliability and objectivity studies need to be conducted to test the feasibility of this instrument when used in the field.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation is not simply the culmination of years of study. It closes an important chapter in my life. The pages printed in this work also reflect the relationship with many inspiring people I have met since I started my graduate work at the University of Pittsburgh. The list is long, but I feel the need to acknowledge all those who have contributed to my development as a scholar and teacher:

First, I would like to express the deepest appreciation to my advisor, Dr. Jere Gallagher, a gracious mentor who has been in every step of my graduate work. Without her guidance and persistent help this dissertation would not have been possible.

To my committee members Louis Pingel, Mary Duquin, and Sara Piva for their encouraging words, thoughtful criticism, and time and attention during busy semesters.

To Marilyn Ross and Louis Fabian for providing me with financial support, which allowed me to complete my studies in the U.S.

To my co-researchers Fabio Fontana and Oldemar Mazzardo for their assistance throughout the development of my dissertation.

To my colleagues from Eastern Illinois University for sharing their enthusiasm on my work: Jill Owen, Robert Sipes, Brent Walker, Scott Crawford, Kevin Hussey, Hasan Mavi, Tony Oliver, Brian Pritschet, and Jeff Willardson.

To Gustavo Almeida, Renata Taveira, and Enio Weissmann for proving logistic support in Pittsburgh, PA.

Finally, warmest thanks must go to my parents, Geny Ferreira Furtado and Ovande Furtado, whose support, although from distance, allowed me to write this dissertation.

1.0 INTRODUCTION

Recent studies (Foley et al., 2008; Hume et al., 2008; Mazzardo, 2008; Okely et al., 2001; Okely et al., 2004) have found a positive relationship between the performance of fundamental movement skills (FMS) and habitual/organized physical activity participation among children and adolescents. Fundamental movement skills (e.g., run, walk, kick, throw, jump) are common movement activities having specific movement patterns, which are believe to form the foundation for more advanced and specific sport and non-sport movement activities (Gabbard, 2007). It is possible that when children and youth feel confident in their skills (self concept), they tend to engage in higher levels of physical activity (Gabbard, 2007). Positive relationships have also been found between fundamental movement skill performance and weight status (Mazzardo, 2008; Okely et al., 2004) among young children. Further, research has shown that children who stay active tend to maintain high levels of physical fitness (Baquet, Twisk, Kemper, Van Praagh, & Berthoin, 2006). Together, these studies provide evidence that proficiency in FMS performance during the early primary grades is likely to contribute to increases in habitual and organized physical activity participation, thus preventing unhealthy weight gain among children and adolescents (Gabbard, 2007).

However, being aware of the importance of developing fundamental movement skills is only part of the process of helping students master such skills. Teachers and practitioners working with younger children must conduct regular assessments to gather evidence about the student's level of achievement in FMS development, and make inferences based on that evidence for a variety of purposes (NASPE, 2004). Data gathered on FMS performance can be used to identify students who are in need of intervention; or results can be used for planning instruction, or even evaluating changes over time of individual students or a group of students. One of today's challenges is for teachers to find the time to conduct assessments.

The National Association for Physical Education has been addressing the issue of the amount of time devoted to assessment in Physical Education (NASPE, 2004) advocating for initiatives that value the importance of assessment as a way for enhancing learning. NASPE (2004) includes assessment as an integral part of instruction. It should enhance learning through a connection with instructional practices. In this view, assessment practices go beyond simply assigning grades to students. However, to accomplish this, teachers need assessment tools that are practical to use in school settings. In the case of FMS, one alternative is for teachers to develop their own instruments, commonly known as authentic assessments. One potential disadvantage of authentic assessments is that they pose a threat to validity and reliability since teachers are unlikely to conduct studies to collect evidence for validity and reliability of test scores. Thus, the alternative for teachers willing to assess fundamental movement skill development on a regular basis is to use already validated, yet practical, assessment tools. This would allow for integration of assessment practices with instruction. However, the assessment tools currently available for testing proficiency in fundamental movement skill development lack in practicability.

A wide variety of assessment tools (Gallahue & Donnelly, 2003; Loovis & Ersing, 1979; Seefeldt & Haubenstricker, 1976; Ulrich & Sanford, 2000) already exist with the intent of providing some degree of information regarding levels of children's fundamental movement skill proficiency. However, the main criticism regarding these instruments is the lack of practicability in physical education settings (Burton & Miller, 1998; Zhu & Cole, 1996). These instruments are widely used for research purposes where time is rarely an issue. For example, researchers typically videotape students performing a variety of FMS and later analyze the videotapes. Although it is possible to videotape students performing FMS in school settings for further analysis, such practice is impractical because of the large number of students in a class in addition to the number of classes. This might potentially prevent teachers from conducting these assessments.

Therefore, the purpose of this study is twofold: (1) to develop a new observational rating system to assess fundamental movement skill performance in children 5- to 10- years of age, and (2) to evaluate the accuracy of this new scale by comparing its ratings with a criteria or the experts ratings. It is expected that the results yielded by this study will allow for improvement of the rating scale.

The instrument proposed in this research, the Furtado-Gallagher Computerized Observational Movement Pattern Assessment System - FG-COMPASS), will enable teachers to easily collect, monitor, analyze, and report student's results, thus allowing assessment practices in FMS performance to be more easily integrated into the instructional process in Physical Education. This instrument is unique in that it uses only a few performance criteria (items) to assess FMS development. Other assessment instruments (Gallahue & Donnelly, 2003; Ulrich & Sanford, 2000) devoted to test the quality aspects of fundamental movement skill performance use between three and six performance criteria in their measurement. The accuracy with which the score interpretation of the FG-COMPASS will be equivalent with already validated tests remains to be tested and is the main purpose of this study.

Because an assessment tool cannot be used to accomplish all purposes, the intended uses of the FG-COMPASS are to: (1) monitor individual progress during and/or following instruction, (2) evaluate effectiveness of the instructional program with the intent of adjusting the curriculum in accordance with student needs, and (3) to detect eventual deficits in fundamental movement skill development.

What follows is a review of the literature regarding fundamental movement skill development and assessment. The literature review is divided into three main sections: (1) understanding fundamental movement skill development; (2) assessing and evaluating fundamental movement skill development, and (3) test construction and evaluation.

1.1 UNDERSTANDING FMS DEVELOPMENT

The purpose of this section is to introduce the topic of fundamental movement skills. Sometimes referred as basic or gross motor skills, this category of movement skills constitutes the construct being assessed by the FG-COMPASS. This section is subdivided into two sections. The discussion that follows is an attempt to clarify the confusion with respect to the use of the term "stages" when referring to fundamental movement skills. The second subsection is intended to discuss the phases of motor skill development.

1.1.1 Stages of fundamental movement skills

Traditionally, the term "stages" has been used to describe changes in intratask motor development. The idea of stages emerged naturally from research in other areas on human development - e.g., cognition, moral, personality (Roberton, 1982). Roberton points out that the theoretical approach of stages used in motor development was based mainly on Piaget's stages of cognition and Kohlberg's stages of moral development.

Proponents of the stage approach, or "classical stage theory", to developmental theory argue that there are universal age periods throughout the lifespan that is characterized by unique behaviors (Gallahue & Ozmun, 2002). According to Payne and Isaacs (2008), such behaviors are not evident until a particular stage begins and may not be evident in the same form when the stage ends. Further, these behaviors last for undefined lengths of time yet are invariant. That is, stages are sequential and cannot be reordered, even though one or two stages may be skipped (Gallahue & Ozmun, 2002).

Although it is impossible to deny the contribution of the stage theory to the understanding of the underlying correlates associated with motor skill changes, researchers (Branta, Haubenstricker, & Seefeldt, 1984; Roberton, 1982) have questioned this assumption of <u>irreversibility</u> and <u>invariance</u> in motor development. This notion of strict phases in motor development has been challenged mainly by studies done to investigate how fundamental movement skill develops during the lifespan. For example, there is evidence suggesting that infants fluctuate between stages when learning to manipulate a lever to gain access to a favorite toy (Koslowski & Bruner, 1972; cited in Branta, et al., 1984). Because of this notion of less rigid "stages" or phases, Roberton (1982) suggests the use of the term "steps" to describe intratask motor sequences.

Therefore, even though the term "stages" is used throughout this paper, it is used to imply a less rigid version of the stage theory. That is, the more contemporary version that dismisses the notion of irreversibility and invariance of intratask motor skill sequences. This is done to avoid confusion with the original works that use such a term. Next, a discussion is provided to address the different phases of motor skill development.

1.1.2 Levels of motor skill development

Although humans are born with certain survival skills (reflexes and reactions) that are embedded in the newborn's nervous system, there are other motor behaviors that are less predetermined (Clark, 2007). Clark points out that these are called preadapted *species-typical or phylogenetic* motor behaviors and require environmental support for their appearance. Among these motor behaviors are the fundamental movement skills (e.g., skip, jump, catch, kick, and run). As the child matures, these movements are gradually controlled by the motor area of the cerebral cortex (Rarick, 1982). Fundamental movement skills provide a framework upon which more complex skills develop (Gabbard, 2007). For instance, the mature movement pattern of stationary ball dribble is necessary for the development of more complex skills such as dribbling a ball while moving forward or the lay up in basketball that requires a combination of dribbling a ball, running and jump.

The term "fundamental movement skills" is often used interchangeably with the term "fundamental movement patterns". This is because as the muscles work in groups, seldom acting alone, the resulting movements (involving one or more joints) are noticeably similar from person to person (Rarick, 1982). This has given origin to the term fundamental movement pattern. According to Rarick the "patterned movements are characterized by an ordered and properly times sequence of subroutines which, when viewed in total, give the movement its quality or form (Rarick, 1982, p. 278).

Motor development specialists (Clark, 2007; Gabbard, 2007; Payne & Isaacs, 2008; Sayre & Gallagher, 2001; Wickstrom, 1983) agree that the period of early childhood, especially between the ages of 1 through 7 years, represents a crucial time for acquiring such movement skills. The phases of motor development are generally classified as the reflexive, rudimentary, fundamental, and specialized movement phases. During infancy and young childhood the reflexive and rudimentary phases are observed. It is believed that these two phases of development represent the fundamental building blocks for the next two phases, fundamental movement patterns and specialized movement (Gabbard, 2004). The fundamental movement phase typically includes ages 2- to 7-years and is believed to be one of the most important periods for motor skill development (Gabbard, 2007; Gallahue & Donnelly, 2003). In this period, "children no longer have to rely on rudimentary motor behaviors to locomote, explore, and manipulate their environment" (Payne & Isaacs, 2008, p. 300). This is the period for children to master the various kinds of locomotor and object-control fundamental movement skills.

Locomotor movements consist of any skill in which the body is transported in a horizontal or vertical direction from one point to another (Graham, Holt/Hale, & Parker, 2007). Locomotor skills are sometimes subdivided into basic skills and combination skills (Gallahue & Donnelly, 2003). The term 'basic' refers to skills that have one element, whereas 'combination' skills combine two or more elements. The skill of running is considered 'basic', whereas side slide is included in the combination category. This is because running consists of one single action that happens continuously. On the other hand, side slide requires the combination of a sideways step and hop. Examples of other basic locomotor movements are leap, horizontal jump, and hop. Examples of other combination locomotor skills are galloping and skip.

Besides locomotor, FMS can also be classified under the category of object-control skills. These are skills in which the individual interacts with an object (Gallahue & Donnelly, 2003). Manipulation occurs either upon giving (propulsion) or receiving (absorption) force to or from objects. Throw, kick, and striking are examples of propulsive object-control skills, whereas catch and trapping are examples of absorptive skills.

The preceding discussion addressed the importance of the phase in skill development known as “fundamental movement skills”. Of equal importance is assessing such skills. Perhaps, the biggest concern, especially in school settings, deal with the feasibility of assessing these skills. Several factors might prevent physical educators and practitioners from conducting regular assessment in fundamental movement skill development including, lack of time devoted physical education and the number of students in class. Issues regarding assessment of FMS development are addressed next.

1.2 ASSESSING FMS DEVELOPMENT

In the proceeding sections a discussion will be provided to address issues related to FMS assessment. What follows is a brief overview of the techniques (composite vs. component) used to assess FMS performance. Next, a discussion is provided that addresses the techniques used for assess FMS development.

1.2.1 Techniques for observational assessment of FMS

The first documented attempts to understand how one's motor behavior changes over time can be traced to 1877 with Darwin's biography of his own child as well as a series of studies carried out by the German physiologist Prayer around the same date (Thelen, 2000). In the 1920's and 30's, a number of physicians and psychologist (Gesell, 1929; Shirley, 1931; and McGraw, 1935; in Seefeldt & Haubenstricker, 1982) documented the sequential changes occurring in infancy and early childhood. These early works mainly charted the motor milestones (Kelso & Clark, 1982). At that time, "specific behaviors were recorded according to the chronological age of the subjects and their order of appearance in the movement repertoire" (Branta et al., 1984, p. 468). Although much has been accomplished since 1920, one fundamental question remains, that is, how motor behavior (e.g., fundamental movement skills) changes during the life span.

Inspired by the works of the beginning of the 20th century, more contemporary motor development specialists (Branta et al., 1984; Roberton, 1982, 1989; Roberton, Williams, & Langendorfer, 1980; Seefeldt & Haubenstricker, 1982) made significant contributions to the understanding of the underlying correlates associated with the process of changes in motor behavior, especially, with regards to fundamental movement skill development. Using FMS as the base of investigation, these researchers have attempted to overcome much of the criticism that the early works were based primarily on a framework of simply charting changes, "when" change occurs (Rarick, 1982). According to Rarick, these contemporary researchers focus more on the question of "how" changes in FMS occur. From these studies, two approaches have emerged. These are the *composite approach* and the *component approach* to fundamental movement skill evaluation.

1.2.2 Composite vs. composite

Two distinct, yet not necessarily opposing, approaches are used to describe movement pattern characteristics. One describes changes in the configuration of **body parts** (component). The second describes changes in the configuration of the **total body** (composite). Ulrich and Branta (1988) stated that "while each approach has its strengths and weaknesses, the resultant performance descriptions are more similar than they are different. Each has been used successfully in research and clinical application" (Ulrich & Branta, 1988, p. 203). Despite the differences, proponents of both approaches agree on three crucial issues, that is, there is high variability among individuals with regards to: a) the age at which development of a specific motor skill emerges, b) the speed of development, and, c) the amount time of time necessary to mature (Branta et al., 1984). Both approaches are explained in detail next.

1.2.2.1 The component approach

The component approach is based on the premise that there are sequential changes in the configuration of **body parts** so that, for the same skill, a sequence describing, for example, arm action is differentiated from a sequence describing the leg action (Painter, 1994). The method assigns a step (stage) classification score (step 1 through step 5) for each of the body components involved in the performance of the FMS. Roberton and colleagues (Roberton, 1977, 1982, 1989; Roberton et al., 1980; Runion, Roberton, & Langendorfer, 2003) have provided important contribution to the development of this approach.

The approach was introduced to the study of FMS evaluation in 1977 when Roberton used two sets of body component categories to describe the overhand throw for force (one for arm action and the other for pelvic-spinal action). The findings pointed to the premise that development of the two components appeared to occur at different rates. Subsequent to the 1977 study, other studies were conducted for different skills including punting (Roberton & Halverson, 1984); sidearm striking (Harper & Struna, 1973); hop (Halverson & Williams, 1985); and the standing long jump (Clark & Phillips, 1985).

1.2.2.2 The composite approach

Unlike the component approach, the composite approach evaluates the body as a whole. This method assigns an overall stage classification score (stage 1 through stage 5). Thus, body configuration for each stage describes the movements of arms, legs, trunk, and head for a given level of performance (Painter, 1994). Much of the research supporting the composite approach has been done by Seefeldt and colleagues (Branta et al., 1984; Seefeldt & Haubenstricker, 1982, 1976). Similar to the component approach, there are studies that validate composite assessments of fundamental movement skills including throw, catch, horizontal jump, kick, hop, skip, striking, and running (Seefeldt, Reuschlin & Vogel, 1972, in Haywood & Getchell, 2005).

The fact that in the composite approach body parts are evaluate as a whole led proponents of the component approach to argue that the composite approach may not be adequate for the evaluation of FMS, for it does not allow for analysis of variability in the development of specific body components (Roberton, 1977). Although admitting that all body parts do not develop as a unit (lockstep fashion), the proponents of the composite approach argue that "there is sufficient cohesion among certain characteristics of a pattern to define those as "stages" of development" (Branta et al., 1984, p. 470). This claim has led to a number of studies in the last decades, thus, supporting the evidence that the composite approach may also be used as an alternate approach to investigate changes in fundamental motor skill development in children. Perhaps, the biggest advantage of the composite over the component approach is that the former is more practical for assessing individuals through observation only (i.e., field observation in school settings). This is because the evaluation is based on observation of the individual's body as a whole, as opposed to specific body components (e.g., trunk, arms, and legs). Therefore the observer does not have to focus independently on 3 to 5 body parts separately.

Although proponents of both approaches have valid arguments with regards to its use, the question whether one should pick one approach over the other depends greatly on the intended uses of the test under development. The component approach might be more appropriate in situations that scores are to be used for critical decisions (e.g., clinical diagnosis, research, and placement). However, there are situations in which the goal is to conduct assessment for tracking one's progress in FMS development, or to perform quick screening tests. In such cases the composite approach might be better suited. This is not to say that tests developed under the component approach should not be used for teaching purposes. It all depends on the amount of time and resources available for the assessments to be conducted.

In short, both the composite and component methods have been used successfully in observational assessment with children. Presented next is a discussion regarding the composite 3-stage approach.

1.2.3 The composite 3-stage approach

The *composite 3-stage approach* was first proposed by McClenaghan (McClenaghan, 1976 cited in Gallahue & Ozmun, 2002). Perhaps, the biggest advantage of this approach is the fact that it limits the choices to three stages. The simple fact that it only uses three stages of classification might make this method more appealing for use in school settings where assessment of fundamental movement skills is usually accomplished through observation. The work of

McClenaghan resulted in an assessment instrument, which was published by McClenaghan and Gallahue (1978) under the name of Fundamental Movement Pattern Assessment Instrument (FMPAI).

The FMPAI is an informal assessment tool that can be used to classify individuals at the 'initial', 'elementary', or 'mature" level with respect to fundamental movement skill development. The instrument is best used to assess movement changes over time since it compares student results to pre-established criteria rather than group norms. Gallahue and Donnelly (2003) state that the instrument has high reliability among trained observers.

The first version of the instrument included only five fundamental movement skills (running, horizontal jump, throwing, catch, and kick). The developmental sequence for each of these five skills was based on the review of the biomechanical literature (Gallahue & Ozmun, 2002). Subsequently, the test was expanded to include several other assessment tasks. These additional skills were walking, vertical jumping, hop, galloping, slide, striking, body rolling, dodging, and one-foot balance.

Recently, a newer version of the test was published (Gallahue & Donnelly, 2007). In the last version the authors added a second battery of tests following a component approach. The composite assessment helps understanding the general picture of the group's level of ability and to identify the children experiencing difficulty; however, a second assessment using the component assessment allows the teacher to pinpoint exactly where the problem lies (Gallahue & Donnelly, 2007).

Although less demanding than previous approaches, observational assessment of FMS development using the composite 3-stage approach is still a daunting task for teachers. The number of performance criteria for the remaining skills used in the FMPAI ranges from 4 to 8.

The solution for this problem is to decrease the number of performance criteria for each *stage classification score* thus making the assessment tool easier to be administered. This would allow test administrators to focus on a fewer number of performance criteria, which could make the process of observational assessment more practical. Haywood and Getchell (2009) have attempted to do that in the last edition of their textbook *Life Span Motor Development*.

1.2.4 The observation plan approach

Haywood and Getchell (2009) have provided an innovative method to testing FMS development based on skill performance observation. The authors refer to this technique as the *observation plan approach*. The authors selected only a few, yet important, performance criteria from the original developmental motor sequences and constructed a decision tree, thus allowing observers to make quick judgments on the development level of a particular individual by completing a quick "yes" or "no" for checkpoints (Haywood & Getchell, 2009). The notion of simplifying DMS is not new.

Taylor (1979) called for the importance of selecting only those features from the DMS that are crucial to movement efficiency. Further, Painter (1994) states that "the distinctly observable behaviors [of DMS] should be differentiated from the less observed behaviors, and the range of movement behaviors should be limited to facilitate observation" (Painter, 1994, p. 9). Also, Painter has called for research to determine how to modify the originally hypothesized DMS into more functional assessment instruments for practitioners.

The method proposed by Haywood and Getchell (2009) is an important step toward the development of assessment tools that are easy to use in field assessments. Inspired by their innovative approach (observation plans approach), the developers of the FG-COMPASS decided to extend their work combining the observational plan approach (Haywood & Getchell, 2009) with the composite 3-stage approach (McClenaghan & Gallahue, 1978). The result of this combination is being referred to as *composite decision-trees approach (CDTA)*. The CDTA combines the idea of limiting the number of the stage classification scores to only three (e.g., initial, elementary, mature), and selecting only key performance criteria for the assessment tasks. In addition, the CDTA is based on composite skill analysis, not component skill analysis, which is believed to facilitate FMS assessment through observation. This modified approach is explained next.

1.2.5 Composite decision tree approach

A new method of assessing FMS development is being proposed in this study that uses only three performance criteria as part of the assessment tasks. In addition, the assessment tasks, hereafter called *composite decision trees* (CDTs), are presented in the form of a horizontal decision tree to facilitate assessment (see Figure 1). Although three performance criteria are used in the CDT below, only two are actually used in the decision process. Also, the method used by the developers of the current test differs from that of Haywood and Getchell's (2009) in that it uses the composite approach to FMS evaluation. By relying on the composite, instead of the component approach, the developers of the current assessment tool seek to provide a practical assessment that can be easily used by teachers in school settings.

The adequacy of the composite decision trees is critical for the current test. In order for score interpretation of the FG-COMPASS to be meaningful, the CDT must accurately classify individuals into their actual membership groups. What follows is a discussion of how each composite decision tree for each assessment task was developed.

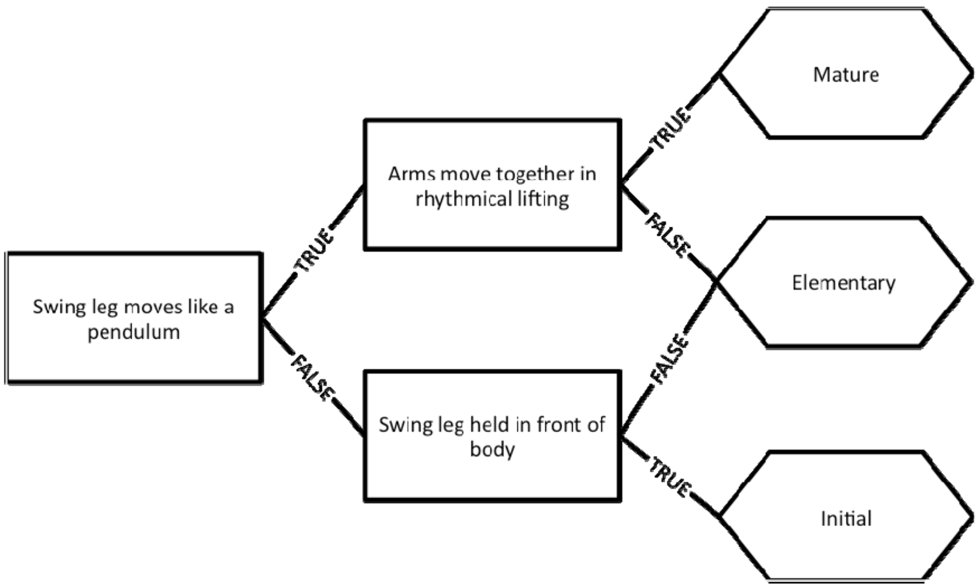

Figure 1: Composite decision tree for hop

1.2.6 The development of the composite decision trees

The *composite decision tree schema* shown in Figure 2 provides an explanation as to how the CDTs for each assessment task was developed. A 'tree' is always read from the left to the right hand side and includes three levels (discriminatory-decision, confirmatory-decision, outcome-decision) and six nodes (three decision-nodes and three outcome-nodes). Each level and each node within the levels have specific goals within the 'tree'. By convention, the person who uses a composite decision tree to assess individuals is referred to as the observer.

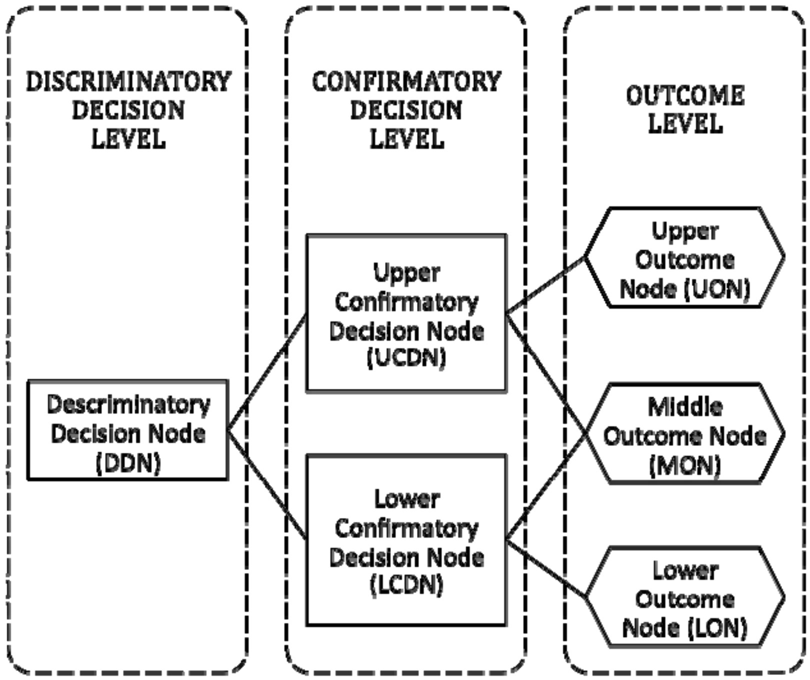

Figure 2: Composite decision tree schema

The first level of the tree is called *discriminatory-decision level* (DDL). This level holds a single decision-node called discriminatory decision-node (DDN). The DDN holds a performance criterion that works as a strong discriminator within the 'tree'. That is, the performance criterion selected for the DDN must be a strong discriminator. In other words, the observer using the composite decision tree should, from the beginning, be able to differentiate the individual who is being assessed whether he/she is at one of the two most extreme levels of FMS development (initial or mature).

The second level of the decision tree is called *confirmatory-decision level* (CDL). This level holds two different nodes, that is, the upper confirmatory-decision node (UCDN) and the lower confirmatory-decision node (LCDN). The UCDN holds a performance criterion that works as a confirmatory decider within the decision tree. The purpose of the UCDN is to confirm that the examinee is indeed at level 3 (mature). On the other hand, the LCDN holds a performance criterion that also works as a confirmatory decider, but its purpose is to confirm that the examinee is indeed at the level 1 (Initial Level). If either the UCDN or LCDN fail to confirm their predicted skill level, the outcome decision will be for level 2 (elementary level).

The third level of the tree is called the outcome level (OL). This level holds three different nodes. These are the upper outcome node (UON), the middle outcome node (MON), and the lower outcome node (LON). These nodes simply hold the final decision that is reached by the system, which may be either initial level, intermediate level, or mature level. An example will help to clarify the concept of using the schema above to develop the composite decision trees. Consider the composite decision tree for overhand throw in Figure 3.

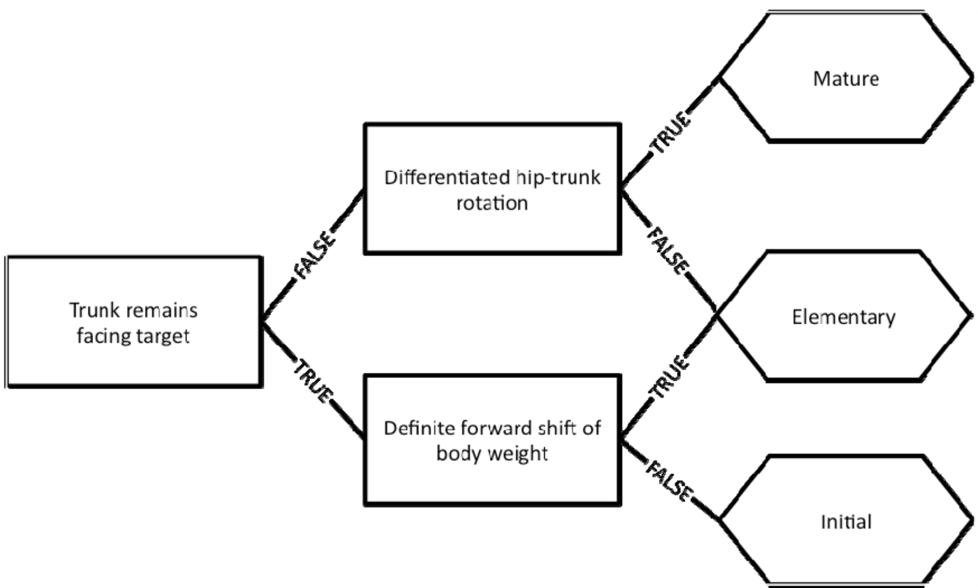

Figure 3: Composite decision-tree for overhand throw

The first step for the construction of the tree depicted in Figure 3 was to decide which performance criterion should be placed at the discriminatory-decision node. Recall that according to the convention adopted earlier, the performance criterion at this level must be strong enough to discriminate between the two most extreme levels (initial and mature). The performance criterion, *trunk remains facing target*, is believed to be such a strong discriminator since individuals performing the overhand throw at the mature level do not show such a pattern, at all. Based on the 3-stage composite approach for overhand throw (Gallahue & Donnelly, 2003), a child who consistently shows the pattern of keeping his/her trunk facing the target during the execution of the overhand throw is considered at the initial level. However, such a decision is still partial since, according to the convention used by the composite decision tree approach, it needs to be confirmed in a subsequent step. This is done at the second level of the tree.

The second step in the development of the CDT for the overhand throw is to select two confirmatory performance criteria. One performance criterion is to be placed at the upper confirmatory decision node and the second at the lower confirmatory-decision node. Thus, if the response for the performance criterion located at the discriminatory level (trunk remains facing target) is “FALSE”, then the performance criterion located at the upper confirmatory decision node (differentiated hip-trunk rotation) should confirm that the examinee being assessed is indeed at the mature level. However, if it fails to confirm, then the examinee is probably at the elementary level. The same applies for the lower confirmatory-decision node. If the response for the performance criterion located at the discriminatory level is “TRUE”, then the performance criterion, *definite forward shift of body weight*, should confirm that the examinee is at the initial level. If it fails to confirm, then the examinee is at the elementary level. As discussed in the beginning of this section, the construction of the decision trees followed a plan of development to ensure the most critical performance criteria were selected. This section concludes the discussion regarding assessing and evaluating FMS. What follows is a brief introduction to the next main topic discussed in this review, that is, test construction and evaluation.

1.3 TEST CONSTRUCTION AND EVALUATION

This section deals with issues related to test construction and evaluation. It starts by describing the two approaches used in test development. Then, the construct being assessed by the FG-COMPASS is briefly discussed. Finally, the definition of validity is provided followed by a discussion of the techniques used in estimating validity.

According to the American Educational Research Association (AERA, 1999), "test development is the process of producing a measure of some aspect of an individual's knowledge, ability, interests, attitudes, or other characteristics by developing items and combining them to form a test, according to a specific plan" (AERA, p. 37). The construction of an assessment tool is a long process that involves several steps. Although each assessment tool is unique, the process of development follows a general format. The first step in developing an assessment tool is the decision about the approach under which the test will be developed. Two approaches are commonly used when developing assessment tools.

1.3.1 Norm and criterion-referenced approaches to test development

Tests within the field of physical education are developed under either the Norm-Referenced Approach (NRA) or Criterion-Reference Approach - CRA (Burton, 1998). Tests developed under NRA and CRA differ in terms of their purposes and thus allow for different interpretation of student performance.

There are occasions in which teachers want to compare students' performances (between-individual comparison). Individual differences are anticipated since some students are expected to perform better than others on a given behavior (Safrit & Wood, 1995). The score is compared to a set of norms, which are developed and provided by the developer of the test. For example, scores on the TGMD-2 (Ulrich & Sanford, 2000), a norm-referenced test designed to test FMS performance, are usually compared with the test's national norms for the individual's gender and age group. This allows the examinees to be compared to each other on the basis of the norm. Often, though, teachers are not concerned with individual differences, but rather how an individual compare with a standard that everyone is expected to meet. In such a case, a test developed under the criterion-referenced approach is preferred.

Sometimes called a mastery test, a criterion-referenced test is referred as an assessment instrument with a predefined standard of performance and with the standard tied to a specific domain of behavior (Safrit & Wood, 1995). Because tests developed under CRA compare students to a standard of performance, they use a specific terminology, which reflects that purpose. Each scoring category indicates how the student did in relation to the standard (e.g., advanced/proficient/basic; master/nonmaster; initial/elementary/mature, etc.). Further, because the main focus of CRA is on *what* test takers can do and what they *know*, not on how they compare with others, scores are easily used for tracking changes over time (Anastasi & Urbina, 1997).

The current assessment tool is being developed under the criterion-referenced approach. Therefore, the current test is characterized mainly for use to track individuals' performance over time rather than determine individual comparisons. Next, validity issues are discussed.

1.3.2 Validity for criterion-referenced tests

In developing criterion-referenced assessment tools, validity becomes an important psychometric concern. Validity is the extent to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (AERA, 1999). Therefore, validity has to do with the clarification and justification of the intended uses and interpretations of observed scores (Kane, 2001). Although the same definition of validity for a norm-referenced test applies to a criterion-referenced test, the techniques for estimating validity are different for these two approaches. (Baumgartner, Mahar, Jackson, & Rowe, 2007). The two procedures commonly used to estimating validity for criterion-referenced tests are domain-referenced validity and decision validity (Barrow, McGee, Tritschler, & Barrow, 1989; Baumgartner et al., 2007; Safrit & Wood, 1995). Although both sources of validity are explained below, this study is focused on collecting evidence for *decision validity* only since initial evidence for *domain-referenced validity* was collected in a previous study (Furtado, Jr., 2004).

1.3.2.1 Domain-referenced validity

Domain-referenced validity is used to collect evidence for the adequacy of the test as a measure of the criterion behavior (Safrit & Wood, 1995). In this context, the question being asked is whether the items proposed by the test developer do, in fact, constitute a representative sample of the wider domain about which to make inferences (Thorn & Deitz, 1989). In addition, items are judged based on their importance (Anastasi & Urbina, 1997). Importance/relevance focuses on whether assessment-tasks or items are included in the test user's domain, which is defined by the test developer (Nitko, 2001). According to Safrit (1995), domain-referenced validity has many similarities with both content validity and logical validity, which are two procedures used to estimate validity of tests being developed under the norm-referenced approach. Initial evidence for domain-referenced validity for the FG-COMPASS was collected in a previous study (Furtado, Jr., 2004), which is discussed next.

Twenty content experts (Kinesiology professors = 8; and experienced PE teachers = 12) served as subjects for the study. An Internet-based item review form was used to collect information from experts. Experts rated content at the test level[1] and item level[2]. At the test level, content was rated for seven different questions addressing the match between the test content and the general test characteristics (e.g., match between the proposed set of items and the test purpose). This set of questions was rated on a five-point likert-type scale with 1 being poor and 5 being excellent. Experts also rated content at the item level (31 items divided in 7 categories) using a four-point likert-type scale where 1 was not important at all and 4 was very important. Descriptive statistics (percentage of responses and median) were used along with qualitative procedures for data analysis. The analysis of content at the item level yielded revision of eleven items. Four items were included based on experts written comments. Finally, two items were dropped. The analysis of the content at the test level showed that questions were rated as very good or excellent by 80% or more of the judges. As a result of this study, eleven assessment tasks (hop, horizontal jump, leap, side slide, skip, batting, catch, hand dribble, kick, overhand throw, and strike) were selected for further analysis.

The study described above was carried out with the intent to collect evidence for the adequacy of the FG-COMPASS as a measure of the criterion behavior (domain-referenced validity). Because the effectiveness of this test lies mainly on the accuracy of its measurement (composite decision trees) to classify individuals according to their developmental level (FMS), the accuracy of classification must be determined. This is done thorough the collection of decision validity evidence.

---

[1] Test level: general questions about the test.
[2] Item level: specific questions about each item.

1.3.2.2 Decision validity

Another procedure for estimating validity with a criterion-referenced test is called *decision validity* (Barrow et al., 1989; Baumgartner et al., 2007; Safrit & Wood, 1995). Test scores are sometimes used to allocate individuals to different categories based on their levels of performance on a given construct. In this context, evidence is needed to judge the suitability of using a test when classifying or assigning a person to one category versus a second or third category (AERA, 1999). The fundamental question in such cases is: *How accurately do test scores predict criterion performance?* To answer this question, test developers need to collect evidence for decision validity. Thus, decision validity demonstrates that the assessment instrument can accurately classify individuals according to their actual group membership (e.g., initial, elementary and mature).

Overall, the purpose of this study was to develop a new observational rating scale to assess fundamental movement skill development of children 5- to 10- years of age. In addition, initial validity evidence was collected to support the accuracy of the rating scale. The ratings of undergraduate students were compared with that of a criterion (expert ratings). The main question answered in this study was whether the newly developed rating scale, which uses only three performance criteria, could be used differentiate individuals of different skill levels.

2.0 METHODS

This study is better organized by dividing this section into Phase I and Phase II. During Phase I video clips of children performing locomotor and object-control fundamental movement skills were developed. These videos were subsequently used in Phase II to evaluate the efficacy of the composite decision trees. Each phase is explained in detail next.

2.1 PHASE I: DEVELOPMENT OF THE VIDEO CLPS

First, children of different age groups were videotaped performing the eleven fundamental movement skills that comprise both subtests (locomotor and object-control) of the FG-COMPASS. Then, the film material was edited and individual video clips were created. Each video clip, which consisted of a child performing one of the eleven fundamental movement skills, was classified based on the 3-stage composite method (Gallahue & Donnelly, 2007).

2.1.1 Videotaping the subjects

One hundred and thirty-three children ranging in age from 6 to 11 years were videotaped performing each of the eleven locomotor and object-control fundamental movement skills. The children were volunteers from a private K-8 school located in Pittsburgh, PA. Parental consent for videotaping was obtained (see APPENDIX B). Children were taken from their regular physical education classes five at a time. Verbal instructions and filming procedures were standardized. Due to time constraints, not all children were videotaped performing all eleven fundamental movement skills. A Sony HDR-HC7 high definition camcorder with 10x optical zoom (2.7'') was used to film the children. Children were given five to six trials on each assessment task. This was done to ensure that enough data would be available for further analysis. Overall, the filming produced approximately 300 minutes of video (five 60-minute Sony MiniDV tapes). The following is an explanation of how each FMS performance captured on tape was used to produce both training and testing video material to be use in Phase II of this study.

2.1.2 Editing and classification of the video clips

Performances captured on tape were further edited and separated by skill (e.g., skip, kick, batting, strike). Then, each performance was classified by skill level (initial, elementary, or mature). Both procedures are explained next.

First, individual performances (video clips) were extracted from the videotapes and separated by skill level. This was done by transferring the content of each tape to an Apple MacBook Pro laptop (2.5Ghz. Intel Core 2 Duo). The software iMovie HD™ was used for editing the video clips. The editing process yielded 446 individual video clips (hop = 29, horizontal jump = 36, leap = 60, side slide = 54, skip = 41, batting = 32, catch = 41, hand dribble = 50, kick = 24, overhand throw = 45, and strike = 34) that were identified for further analysis. Each video clip depicts a child performing one of the eleven fundamental movement skills selected for the FG-COMPASS. With the exception of hand dribble, all video clips depict a child performing three trials for each performance task. For the hand dribble*,* students were asked to dribble a ball (stationary) for about 20 seconds. Each edited video clip was classified into one of the three developmental levels (initial, elementary or mature). The composite 3-stage approach (Gallahue & Donnelly, 2007) was used to classify each video clip according to its developmental level. The principal investigator of this study did the classification of all 446 video clips.

Next, data was reduced by selecting three video clips for each of the three skill levels for each fundamental movement skill. This resulted in 99 video clips. A second rater was asked to classify these 99 video clips. This was done to ensure accuracy of the classifications. In case of disagreement between the two raters on a given video, a third rater was asked to classify that video. Then, all three raters watched the video together so that a consensus could be reached with regards to the developmental level of the video. Whenever a consensus could not be reached among all three raters, the video was replaced by a different video that the three raters agreed upon it. Thus, an updated list of 99 videos was generated. These videos were used for testing purposes during the Phase II of this study. The remainder 347 video clips were used for training purposes.

All of the developmental levels for each skill were represented in the sample. An effort was made to ensure representation of all ages within each skill, but the distribution was not equivalent. This is partially due to the fact that only three videos comprised each level for each skill. When comparing the age (months) distribution in each skill, leap had older children (*M*= 117, *SD*= 19), followed by batting (*M*= 116, *SD*= 20), horizontal jump (*M*= 114, *SD*= 30), strike (*M*= 109, *SD*= 20), side slide (*M*= 106, *SD*= 28), hop (*M*= 103, *SD*= 25), hand dribble (*M*= 101, *SD*= 22), skip (*M*= 101, *SD*= 25), kick (*M*= 98, *SD*= 27), overhand throw (*M*= 98, *SD*= 15), and catch (*M*= 85, *SD*= 15). Appendix D shows the descriptive statistics for age distribution across each skill and skill levels. In addition, with the exception of catch and strike*,* all skills had three video clips representing each level. The initial level of catch was represented by two videos. To keep the total number of videos consistent among skills (e.g., nine videos) a fourth elementary-level video was added to the skill of catch. The same is true for the striking. The Phase II is explained next.

2.2 PHASE II: DEVELOPMENT AND EVALUATION OF THE DECISION TREES

During the Phase II the rating system (composite decision trees) for each skill was developed. In addition, information was gathered to determine whether teachers could use the composite decision trees to evaluate the skill levels of children. Phase II is explained in detail next.

2.2.1 Subject selection

Potential participants either called or emailed in response to an advertisement posted on campus and outside of the university environment. The participants were pre-screened via telephone (see APPENDIX E for the telephone script). This was done to ensure participation eligibility. To participate in this study, subjects needed to either be enrolled in or have graduated from a graduate/undergraduate K-12 Physical Education Teacher Certification program. The sample comprised of thirty subjects (21 males and 9 females). Subjects were two (7%) freshman, two (7%) sophomores, thirteen (43%) juniors, and thirteen (43%) seniors. Fifty-three per cent of the subjects reported no experience teaching physical education. Twenty-seven per cent of the subjects reported having taught physical education for less than 6 months, and 13% for more than six months but less than one year. Only two subjects taught structured physical education for more than one but less than two years.

Subjects received a total of $40 for participating in this research study. In addition, two subjects received an additional $50 based on their ratings. This was done by comparing percentage of agreement between the subject's responses for each video clip with its respective criteria (correct response). The two best-ranked subjects received the extra money. This was done to motivate the participants to correctly classify each video clip according to its actual developmental level.

2.2.2 Training session

Subjects were required to undergo a training session prior to the testing session. The training session, which lasted approximately 1 hour, was done individually and took place 3-5 days before the testing session. During the training session subjects were asked to complete eleven computer-based training modules; one module for each fundamental movement skill proposed for the FG-COMPASS.

Each module included three video clips, one for each skill level (initial, elementary, and mature). A demonstration was developed to help subjects to get familiar with the computer-based training tool. Provided next is an explanation that shows how subjects completed each training module. Screenshots from the computer-based training tool are used to help the explanation.

Figure 4 shows the first screen presented to subjects during the training session. Upon selecting a given module subjects were presented with the *visual cues* screen (Figure 5). This screen contained visual cues (see APPENDIX H) to which subjects were encouraged to focus their attention while watching the video clips. For instance, while watching the videos for the skill of catch, subjects were asked to focus their attention to "arms" and "hand grasp" during the action. The concept of the "visual cues" was developed to help subjects focus on specific aspects of the performance. In addition to presenting the visual cues to the subjects, physical demonstrations were provided emphasizing the key aspects of the skill being presented. This procedure has been suggested by Painter (1994) to ensure subjects are clear with regards to the trait being assessed.

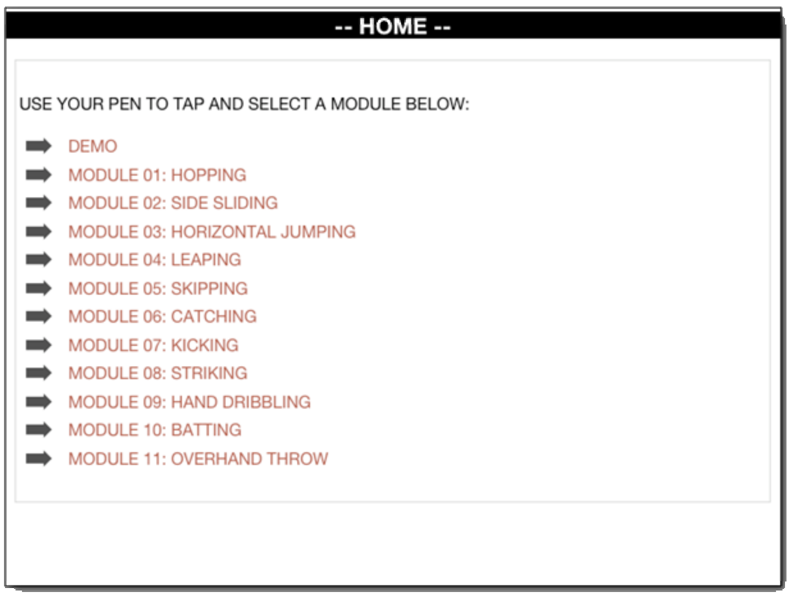

Figure 4: Home for the computer-based training tool

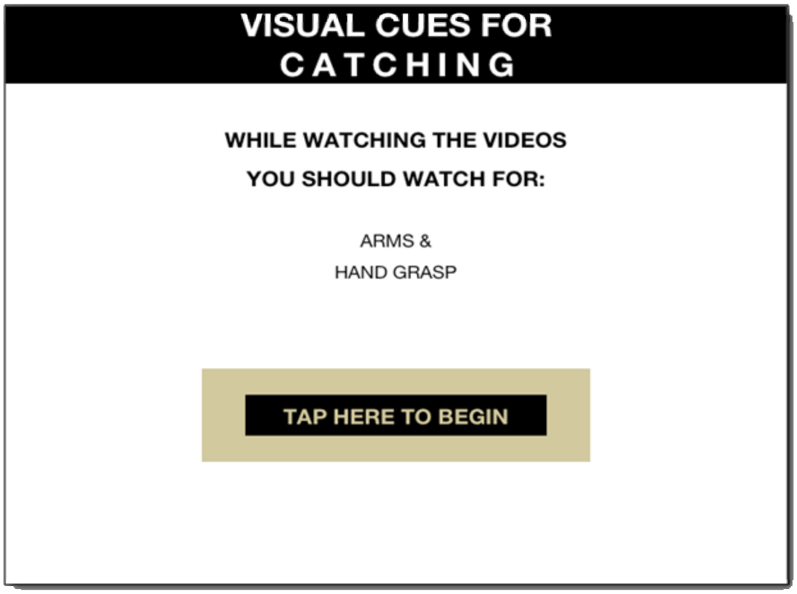

Figure 5: Visual Cues screen for the computer-based training tool

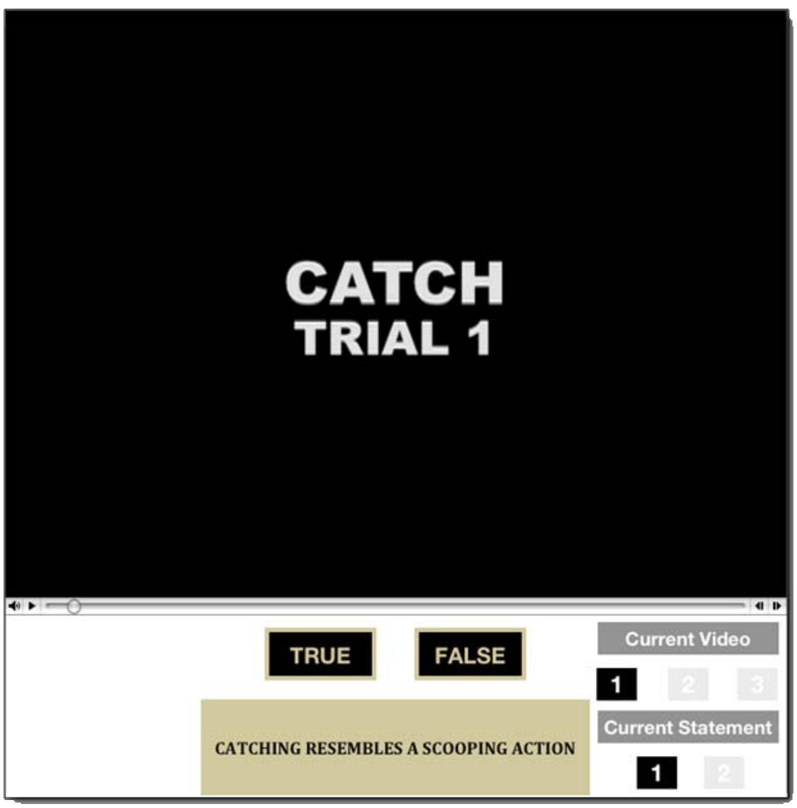

Figure 6: Statement screen for the computer-based training tool

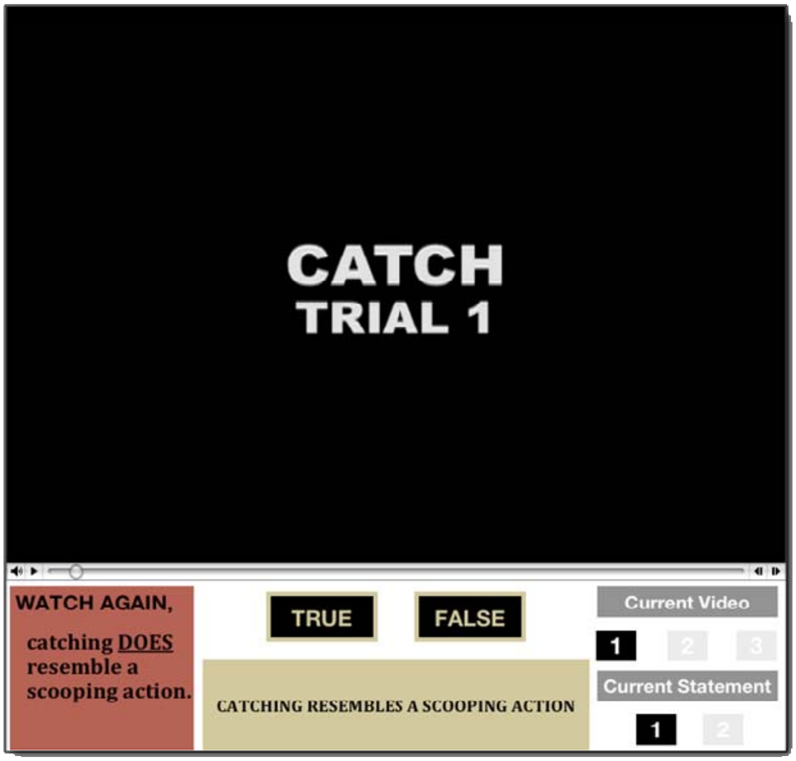

Figure 7: Feedback screen for the computer-based training tool

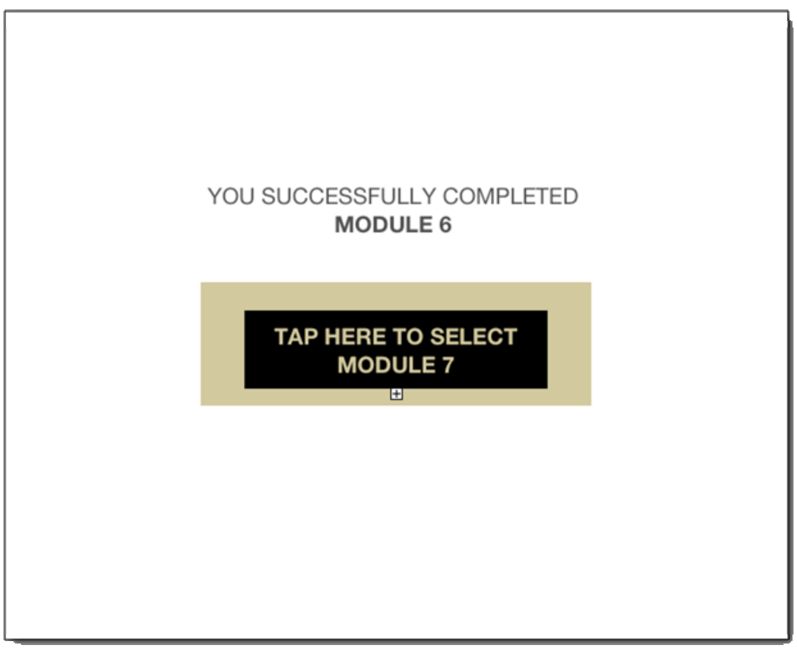

Figure 8: Completion screen for the computer-based training tool

After studying visual cues, subjects then proceeded by clicking the “tap here to begin” button. In the next screen (Figure 6), subjects were presented with the first video along with the first statement. At this point, subjects were instructed to play the video then answer “true” or “false” to the first statement. In case of a correct response; that is, the response was in accordance with the development level of the video, statement two was presented. Subjects then would proceed the same way answering “true” or “false” to the second statement. However, if the response to the first statement was incorrect, then the feedback screen (Figure 7) was presented. Whenever presented with this screen, subjects were instructed to read the feedback (red square located in the left lower corner of the screen) and watch the video again. After watching the video for the second time, subjects were encouraged to ask questions; otherwise they were to change their original response and then proceed to the next video. This procedure continued until subjects completed rating all three videos. At the end of the third video a confirmation screen was presented (Figure 8). This sequence continued until the subjects completed all 11 modules. At the end of the eleventh module, subjects were asked whether they wanted to revisit any of the modules for further clarification. Otherwise, they were told the training session was over. In addition, any unusual comment or questions about the statements or the videos were written down during the training session.

2.2.3 Testing session

The testing session was done individually and carried out 3-5 days after the testing session. Each subject rated ten video clips for each FMS (one practice and nine testing video clips). Thus, each subject rated 110 video clips. Subjects were seated 10.5 feet from a 33 x 18 inches white screen where the video clips were projected. In front of the subject was a laptop computer with touch screen capability that was used to displaying the questions for each video clip. Subjects rated the level of each child appearing on the video clips by answering "true" or "false" (tapping the computer screen) to two statements prompted on the computer screen.

To control for rater fatigue, subjects took two 5-minute breaks. The first break was after the completion of the third assessment task (30 video clips), and then again after the completion of the seventh assessment task (70 video clips). Both the order of the assessment tasks and the order of video clips within each assessment task were randomized for each subject. See Appendix C for randomization order. Also, during testing, subjects were instructed not to answer statements based on the apparent age of the children as skill level was distributed across all age levels for most of the assessment tasks. Appendix D shows the descriptive statistics for age distribution across each skill and skill levels.

One last note is that subjects were not told how many videos of each level they would be rating. That was done to prevent subjects from engaging in guessing during the classification of the videos. Besides, subjects received feedback about the correct classification during the training, but not during the testing session.

2.3 DATA ANALYSIS

The analysis used in this study tested the effectiveness of the composite decision trees in evaluating the skill level of the children. To do this, contingency tables were first generated. Then, the proportion of overall agreement (A_o), and the weighted kappa (K_w) were calculated for each skill. Finally, the proportion of agreement (P_s) and the percentage of agreement were calculated for each category (initial, elementary, and mature) within a skill. Each procedure is explained in further detail next.

2.3.1 Observed agreement

The observed agreement (A_o) is the proportion of cases (videos) for which the expected and observed scores agree. The expected values are the actual skill level of the children observed. This is considered the actual level because a panel of three experts agreed upon the level of each child. The observed scores were the responses given by the subjects when rating the skill level of each child. Therefore, the actual skill level of each child was compared with its observed score given by the subjects of this study.

The formula for calculating A_o is presented in Equation 1 below. Equation 1 was generated from the information provided in Table 1. The observed agreement is the sum along the diagonal in the table below divided by the total number of cases.

Equation 1: Formula used to calculate the proportion of overall agreement for each skill

$$Ao = \frac{a+e+i}{a+b+c+d+e+f+g+h+i} = \frac{a+e+i}{N}$$

Table 1: Hypothetical data for expected and observed agreement on three categories

	Expected			
Observed	Category 1	Category 2	Category 3	Total
Category 1	a(.75)	b(.01)	c(.04)	a+b+c(.80)
Category 2	d(.05)	e(.04)	f(.01)	d+e+f(.10)
Category 3	g(0)	h(0)	i(.10)	g+h+i(.10)
Total	a+d+g(.80)	b+e+h(.05)	c+f+i(.15)	N(100)

Note. Values enclosed in parentheses denote hypothetical data in proportion.

2.3.2 Weighted kappa

Despite its popularity, the A_o can, in some situations, be misleading. This is because a certain amount of agreement is expected to occur by chance (Fleiss, 1973). To account for this, the kappa inter-rater agreement statistics can be used. For the present study, the weighted kappa (K_w) was used. The weighted kappa is a generalization of the kappa statistics (Fleiss & Cohen, 1973). It provides a measure of agreement between two raters who classify observations into one of several categories. It allows one to assign different penalties to different mismatch among the classifications. The weighted kappa was calculated using the computer package SAS (9.0).

2.3.3 Proportion of specific agreement

Both the A_o and K_w combine the agreement for each of the categories. However, agreement may differ in each category. To verify the degree of agreement in each category separately, the proportion of specific agreement (P_s) can be calculated. This is done by collapsing each 3x3 table (see Table 2) into a 2x2 table (Spitzer & Fleiss, 1974). This procedure is explained next. To exemplify, Category 2 will be collapsed. First, the average proportion of all subjects by using the marginal totals is calculated. For the hypothetical data in Table 1, this proportion is ½ (.10+.05) = .075. The next step is to find the proportion agreed upon the selected category, which in the presented example is Category 2. The final step is to take the ratio of the two proportions above. The resulting value is the *proportion of specific agreement*. For the hypothetical data on Table 1, the resulting value is (.04/.075)= .53.

Table 2: Hypothetical data for expected/observed agreement on category two

	Expected		
Observed	Category 2	Other*	Total
Category 2	e(.04)	f+d(.06)	e+f+d(.10)
Other*	b+h(.01)	e+a+i(.89)	b+h+e+a+i(.90)
Total	e+b+h(.05)	f+d+e+a+i(.95)	N(100)

Note. Values enclosed in parentheses denote hypothetical data in proportions. Asterisks denote the new category that was formed after category 1 and 3 were collapsed.

2.3.4 Mean percentage agreement/disagreement

Independent of the proportion of specific agreement index, the *mean percentage agreement* for each dimension of the decision tree can be reported. Thus, the mean percentage agreement can be reported for the left or right dimensions of the decision tree or even for the upper and lower levels of the right dimension separately. The left dimension of the decision tree is where the discriminatory statement (DS) is placed. The right dimension is where the two confirmatory statements (CS-upper or CS-lower) are placed. This allows for a closer inspection of the discrepancies noticed with the previously discussed indexes.

3.0 RESULTS

As suggested by Hripcsak and Heitjan (2002) the best approach to study rater agreement depends on the purpose of the study. The ultimate purpose of this study is to improve the FG-COMPASS rating system (composite decision trees). Thus, an approach that gives a degree of agreement among raters, but also that separates the components of agreement is critical.

First, the computer package SPSS (17.0) was used to generate 3x3 contingency tables (see Table 3) for the observed vs. expected scores for each of the eleven skills. The expected values (criteria or expert ratings) comprised the columns of the table whereas the observed values (subjects' responses) comprised the rows. For each table, the count and proportion values for each cell, as well as the marginal totals and grand total are provided. Then, three indexes were calculated from the contingency tables.

Table 3: Observed vs. expected contingency tables for each skill

Skill			Expected			
			Initial	Elementary	Mature	Total
Hop	Observed	Initial	88(32.59)	10(3.70)	1(0.37)	99(36.67)
		Elementary	2(0.74)	64(.23.70)	6(2.22)	72(26.67)
		Mature	0(0.00)	16(5.93)	83(30.74)	99(36.67)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)
Side Slide	Observed	Initial	68(25.19)	15(5.56)	2(0.74)	85(31.48)
		Elementary	18(6.67)	28(10.37)	6(2.22)	52(19.26)
		Mature	4(1.48)	47(.17.41)	82(30.37)	133(49.26)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)
Horizontal jump	Observed	Initial	59(21.85)	33(12.22)	2(0.74)	94(34.81)
		Elementary	31(11.48)	57(21.11)	4(1.48)	92(34.07)
		Mature	0(0.00)	0(0.00)	84(31.11)	84(31.11)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)
Leap	Observed	Initial	70(25.93)	38(14.07)	1(0.37)	109(40.37)
		Elementary	16(5.93)	33(12.22)	13(4.81)	62(22.96)
		Mature	4(1.48)	19(7.04)	76(28.15)	99(36.67)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)
Skip	Observed	Initial	85(31.48)	6(2.22)	0(0.00)	91(33.70)
		Elementary	1(0.37)	75(27.78)	28(10.37)	104(38.52)
		Mature	5(1.85)	9(3.33)	61(22.59)	75(27.78)
		Total	91(33.33)	90(33.33)	89(33.33)	270(100)
Catch	Observed	Initial	50(18.52)	30(11.11)	0(0.00)	80(29.63)
		Elementary	10(3.70)	79(29.26)	11(4.07)	100(37.04)
		Mature	0(0.00)	11(4.07)	79(29.26)	90(33.33)
		Total	60(22.22)	120(44.44)	90(33.33)	270(100)

Table 3 (continued)

Kick	Observed	Initial	18(6.67)	20(7.41)	0(0.00)	38(14.07)
		Elementary	70(25.93)	64(23.70)	9(3.33)	143(52.96)
		Mature	2(0.74)	6(2.22)	81(30.00)	89(32.96)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)
Strike	Observed	Initial	54(20.00)	11(4.07)	0(0.00)	65(24.07)
		Elementary	6(2.22)	106(39.26)	24(8.89)	136(50.37)
		Mature	0(0.00)	3(1.11)	66(24.44)	69(25.56)
		Total	60(22.22)	120(44.44)	90(33.33)	270(100)
Hand dribble	Observed	Initial	60(22.22)	3(1.11)	0(0.00)	63(23.33)
		Elementary	29(10.74)	60(22.22)	5(1.85)	94(34.81)
		Mature	1(0.37)	27(10.00)	85(31.48)	113(41.85)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)
Batting	Observed	Initial	71(26.30)	11(4.07)	2(0.74)	84(31.11)
		Elementary	19(7.04)	68(25.19)	5(1.85)	92(34.07)
		Mature	0(0.00)	11(4.07)	83(30.74)	94(34.81)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)
Overhand Throw	Observed	Initial	87(32.22)	57(21.11)	2(0.74)	146(54.07)
		Elementary	3(1.11)	32(11.85)	2(0.74)	37(13.70)
		Mature	0(0.00)	1(0.37)	86(31.85)	87(32.22)
		Total	90(33.33)	90(33.33)	90(33.33)	270(100)

Table 4 shows a summary of the statistical indexes calculated for the present data. Provided in Table 4 are the values for weighted kappa *(K_w)*, respective confidence limits for K_w, and the proportion overall agreement (observed agreement) for each skill. In addition, the values of the proportion of specific agreement *(P_s)* for each category within each skill are provided.

Table 4: Summary of the different indexes calculated for each skill

	Kappa Statistics		P_s			A_o
Skill	K_w	*95% CI*	*I*	*E*	*M*	
Hop	.85	.81-.90	.93	.79	.88	87%
Side Slide	.61	.54-.68	.77	.40*	.73	66%
Horizontal jump	.70	.64-.77	.65*	.63*	.97	74%
Leap	.61	.55-.68	.70	.43*	.80	66%
Skip	.77	.71-.83	.93	.77	.74	82%
Catch	.72	.66-.78	.71	.72	.88	77%
Kick	.51	.44-.59	.28*	.56*	.91	61%
Strike	.79	.73-.85	.86	.83	.83	84%
Hand dribble	.72	.66-.78	.78	.65*	.84	76%
Batting	.79	.74-.85	.82	.75	.90	82%
Overhand Throw	.74	.68-.80	.74	.50*	.97	76%

Note. I=Initial; E=Elementary; M=Mature. Asterisks denote categories containing videos that were selected for further analysis.

Reviewing the K_w values in Table 4, the agreement was best for the skill of hop (K_w= .85), followed by strike and batting (K_w= .79), skip (K_w= .77), overhand throw (K_w= .74), catch and hand dribble (K_w= .72), and horizontal jump (K_w= .70). The poorest agreement occurred in the skills of kick, (K_w= .51), side slide and leap (K_w= .61). The calculation of the proportion of specific agreement provided relevant information, which allowed for a close inspection regarding the source of the disagreement within each skill. For that, the categories (initial, elementary, mature) with P_s values below .70 were further inspected.

Six skills were selected for further analysis (side slide, horizontal jump, leap, kick, hand dribble, and overhand throw). The next step was to identify the source of the discrepancy within each decision tree identified for further analysis. As discussed in the previous section, the source of disagreement in any decision tree can arise from responses to the first or second statement or both statements combined. In addition, the pattern of disagreement across the videos is worth investigating. If the source of disagreement is from a single video within a category, then there is reason to believe that the source of the discrepancy is associated with that video and not with the decision tree itself. Presented next are the results for each of the six skills identified for further analysis.

3.1 SIDE SLIDE

The proportion of specific agreement was $P_{s(2)}$= .40 for the elementary-level category of side slide, which was considerably low compared to the initial-level ($P_{s(1)}$= .77) and mature-level ($P_{s(3)}$= .73) categories, Thus, further steps were taken to identify where in the decision tree most of the disagreement occurred for the elementary-level category.

The disagreement was not similar across all three videos (Video 4= 36.7%, Video 7= 70%; Video 8= 50%). See Table 5 for further details. The disagreement was greater at the upper confirmatory statement (52.2%), than the lower confirmatory statement (16.7%). Thus the CS-upper did not function as expected when used to classify elementary-level videos.

Table 5: Summary of percentage agreement/disagreement for side slide

	Videos			Mean %	
Category	*% Agreement for 1st Video*	*% Agreement for 2nd Video*	*% Agreement for 3rd Video*	*Agree-ment*	*Disa-greement*
Initial (2, 3, 5)					
DS					
True	23.3**	3.3**	0.0**	-	8.9_a
False	76.7*	96.7*	100.0*	91.1	-
CS-upper					
True	13.3**	0.0**	0.0**	-	4.4_b
False	10.0**	3.3**	0.0**	-	4.4_b
CS-lower					
True	56.7*	86.7*	83.3*	75.6	-
False	20.0**	10.0**	16.7**	-	15.6_c
Elementary (4, 7, 8)					
DS					
True	56.7	90.0	76.7	-	-
False	43.3	10.0	23.3	-	-
CS-upper					
True	36.7**	70.0**	50.0**	-	52.2_d
False	20.0*	20.0*	26.7*	22.2	-
CS-lower					
True	30.0**	6.7**	13.3**	-	16.7_e
False	13.3*	3.3*	10.0*	26.6	-
Mature (1, 6, 9)					
DS					
True	100*	90.0*	100*	96.8	-
False	0.0**	10.0**	0.0**	-	3.3_f
CS-upper					
True	96.7*	80.0*	96.7*	92.2	-
False	3.3**	10.0**	3.3**	-	5.5_g
CS-lower					
True	0.0**	6.7**	0.0**	-	1.1_h
False	0.0**	3.3**	0.0**	-	2.2_h

Note. DS= Discriminatory statement; CS= Confirmatory statement. Single asterisk denotes correct response. Double asterisk denote incorrect response. Numbers enclosed in parentheses denote the actual video numbers for each category. a= disagreement for initial-DS; b= disagreement for initial-CS-upper (adding up both values of *b* amounts the *a* value); c= disagreement for initial-CS-lower (adding up *a* and *c* amounts the total average disagreement for *initial*); d= disagreement for elementary-CS-upper; e= disagreement for elementary-CS-lower (adding up *d* and *e* amounts the total average disagreement for *elementary*); f= disagreement for mature-DS; g= disagreement for mature-CS-upper; h= disagreement for mature-CS-lower (adding up both values of *h* amounts the *f* value) – adding up *f* and *g* amounts the total average disagreement for *mature*.

To better understand this discrepancy, the three elementary-level videos were closely inspected by comparing the performance of the children appearing on the videos with the decision tree used. This qualitative analysis led to the replacement of the CS-upper criterion. The criterion, *action is not choppy and stiff*, was replaced with *clearly airborne throughout action*. Table 6 shows the updated decision tree for the skill of side slide.

Three main reasons led to the replacement of the CS-upper criterion. First, consider the original statement above. It is possible that the double wording at the end of the statement (choppy and stiff) confused some of the subjects. Some of the subjects might have interpreted these as two independent words/behaviors, and for that to be considered false, both had to be present. Second, the "NOT" (double negative) part of the statement might have caused some confusion when subjects were answering "true" or "false" to the statement. Third, even though subjects had to undergo training prior data collection, it could be that they simply could not differentiate between *choppy/stiff* and *not choppy/stiff*. Perhaps selecting a single-meaning and less subjective statement would help increase accuracy of classifications at this level of the decision tree. Therefore, it is believed that the modifications done in the decision tree of side slide will help with classification of elementary level individuals. The new statement, *clearly airborne throughout action*, is expected to be a better discriminator between the initial and elementary levels. Next, the results are discussed for horizontal jump.

Table 6: Final decision tree for side slide

<table>
<tr><td rowspan="4">(DS)
Smooth, rhythmical action</td><td rowspan="2">True</td><td rowspan="2">(CS-upper)
Action is NOT choppy and stiff

Changed to:

Clearly airborne throughout action</td><td>True</td><td>M</td></tr>
<tr><td>False</td><td rowspan="2">E</td></tr>
<tr><td rowspan="2">False</td><td rowspan="2">(CS-lower)
Double hop OR step occurs</td><td>False</td></tr>
<tr><td>True</td><td>I</td></tr>
</table>

Note. M= Mature, E= Elementary, I= Initial. Shaded areas denote the expected path. (DS) Discriminatory statement; (CS-upper) Upper confirmatory statement, (CS-lower) Lower confirmatory statement.

3.2 HORIZONTAL JUMP

The proportion of specific agreement was $P_{s(1)}$= .65 for the initial-level, $P_{s(2)}$= .63 for the elementary-level and $P_{s(3)}$= .97 for the mature-level categories. Next, results are presented for each of the two categories that had a P_s below .70.

The percentage agreement values for the initial-level category were similar for Video 1, 70%; and Video 3, 73.3%, but lower for Video 7, 53.3% (see Table 7). In terms of the location of the disagreement, the right side (CS-lower), 33.3%, had a higher percentage of disagreement compared to the left side (DS), 1.1%. This indicates that the DS functioned as expected when used to classify initial-level videos. The same cannot be said about the CS-lower.

Similar to the initial-level category, the discrepancy observed in the elementary-level category was not similar across all three videos. The disagreement was evident in Video 5 (40%) and Video 6 (53.3%), but not in Video 9 (16.7%). The discrepancy observed in the elementary-level category occurred entirely in the CS-lower (36.7%). All responses taking the upper path in Video 6 (16.7%) and Video 9 (33.3%) were correctly placed under the elementary-level category. This shows that, unlike the CS-lower, the CS-upper functioned as expected when used to classify elementary-level videos in horizontal jump.

Table 7: Summary of percentage agreement/disagreement for horizontal jump

	Videos			*Mean %*	
Category	*% Agreement for 1st Video*	*% Agreement for 2nd Video*	*% Agreement for 3rd Video*	*Agree-ment*	*Disa-greement*
Initial (1, 3, 7)					
DS					
True	0.0**	0.0**	3.3**	-	1.1_a
False	100.0*	100.0*	96.7*	98.9	-
CS-upper					
True	0.0**	0.0**	0.0**	-	0.0_b
False	0.0**	0.0**	3.3**	-	1.1_b
CS-lower					
True	70.0*	73.3*	53.3*	65.7	-
False	30.0**	26.7**	43.3**	-	33.3_c
Elementary (5, 6, 9)					
DS					
True	0.0	16.7	33.3	-	-
False	100.0	83.3	66.7	-	-
CS-upper					
True	0.0**	0.0**	0.0**	-	0.0_d
False	0.0*	16.7*	33.3*	16.7	-
CS-lower					
True	40.0**	53.3**	16.7**	-	36.7_e
False	60.0*	30.0*	50.0*	46.7	-
Mature (2, 4, 8)					
DS					
True	100*	93.3*	90.0*	94.4	-
False	0.0**	6.7**	10.0**	-	5.6_f
CS-upper					
True	100.0*	93.3*	86.7*	93.3	-
False	0.0**	0.0**	3.3**	-	1.1_g
CS-lower					
True	0.0**	6.7**	0.0**	-	2.2_h
False	0.0**	0.0**	10.0**	-	3.3_h

Note. DS= Discriminatory statement; CS= Confirmatory statement. Single asterisk denotes correct response. Double asterisk denote incorrect response. Numbers enclosed in parentheses denote the actual video numbers for each category. a= disagreement for initial-DS; b= disagreement for initial-CS-upper (adding up both values of *b* amounts the *a* value); c= disagreement for initial-CS-lower (adding up *a* and *c* amounts the total average disagreement for *initial*); d= disagreement for elementary-CS-upper; e= disagreement for elementary-CS-lower (adding up *d* and *e* amounts the total average disagreement for *elementary*); f= disagreement for mature-DS; g= disagreement for mature-CS-upper; h= disagreement for mature-CS-lower (adding up both values of *h* amounts the *f* value) – adding up *f* and *g* amounts the total average disagreement for *mature*.

In an attempt to better understand this discrepancy, all six videos (three initial-level and three elementary-level) were closely inspected. This qualitative analysis led to the replacement of the CS-lower. The criterion, *the trunk moves in vertical direction; little emphasis on length of jump*, was replaced with *difficulty in using both feet*. Table 8 shows the updated decision tree for horizontal jump.

The original CS-lower criterion was replaced for two mains reasons. First, when looking at the videos, it was noted that not all children who experienced difficulty in jump far (condition 2), did it with their trunk in a vertical position (condition 1). And because both concepts are part of the same statement, this could have led observers to think that both conditions had to be satisfied in order for the response be considered true. The second reason that led to the replacement of the CS-lower was because it did not function as a strong discriminator. A better discriminator should be for observers to detect whether or not performers are taking off and jump with both feet (new statement), than trying to detect the degree of emphasis on the length of the jump, or whether the trunk moves in the vertical direction. Therefore, it is expected that accuracy of classifications will improve with the replacement of the CS-lower within the decision tree for horizontal jump.

Table 8: Final decision tree for horizontal jump

(DS) Arms thrust forward forcefully on takeoff	True	(CS-upper) Arms move high and to rear during preparatory crouch	True	M
			False	E
	False	(CS-lower) Trunk moves in vertical direction; little emphasis on length of jump **Changed to:** Difficulty in using both feet	False	E
			True	I

Note. M= Mature, E= Elementary, I= Initial. Shaded areas denote the expected path. (DS) Discriminatory statement; (CS-upper) Upper confirmatory statement, (CS-lower) Lower confirmatory statement.

3.3 LEAP

The proportion of specific agreement was $P_{s(2)}$= .43 for the elementary-level category, thus further steps were taken to identify the source of the disagreements.

The discrepancy in Video 1 (56.7%) was similar to of that in Video 9 (60%), but slightly higher in Video 8 (73.3%). Regarding the location of the disagreement, the mean percentage disagreement was greater for the CS-lower (43.2%) than of that for the CS-upper (21.1%). Thus, neither (upper or lower) dimension of the decision tree functioned as expected when used to classify elementary-level videos.

Table 9: Summary of percentage agreement/disagreement for leap

	Videos			Mean %	
Category	*% Agreement for 1st Video*	*% Agreement for 2nd Video*	*% Agreement for 3rd Video*	*Agree-ment*	*Disa-greement*
Initial (1, 4, 5)					
DS					
True	10.0**	13.3**	6.7**	-	10.0_a
False	90.0*	86.7*	93.3*	90.0	-
CS-upper					
True	3.3**	6.7**	3.3**	-	4.4_b
False	6.7**	6.7**	3.3**	-	5.6_b
CS-lower					
True	90.0*	86.7*	73.3*	83.3	-
False	0.0**	10**	20.0**	-	10.0_c
Elementary (2, 8, 9)					
DS					
True	33.3	53.3	13.3	-	-
False	66.7	46.7	86.7	-	-
CS-upper					
True	10.0**	43.3**	10.0**	-	21.1_d
False	23.3*	10.0*	3.3*	12.2	-
CS-lower					
True	46.7**	30.0**	50.0**	-	42.2_e
False	20.0*	17.7*	36.7*	24.8	-
Mature (3, 6, 7)					
DS					
True	96.7*	96.7*	93.3*	95.6	-
False	3.3**	3.3**	6.7**	-	4.4_f
CS-upper					
True	86.7*	76.7*	90.0*	84.4	-
False	10.0**	20.0**	3.3**	-	11.1_g
CS-lower					
True	3.3**	0.0**	0.0**	-	1.1_h
False	0.0**	3.3**	6.7**	-	3.3_h

Note. DS= Discriminatory statement; CS= Confirmatory statement. Single asterisk denotes correct response. Double asterisk denote incorrect response. Numbers enclosed in parentheses denote the actual video numbers for each category. a= disagreement for initial-DS; b= disagreement for initial-CS-upper (adding up both values of *b* amounts the *a* value); c= disagreement for initial-CS-lower (adding up *a* and *c* amounts the total average disagreement for *initial*); d= disagreement for elementary-CS-upper; e= disagreement for elementary-CS-lower (adding up *d* and *e* amounts the total average disagreement for *elementary*); f= disagreement for mature-DS; g= disagreement for mature-CS-upper; h= disagreement for mature-CS-lower (adding up both values of *h* amounts the *f* value) – adding up *f* and *g* amounts the total average disagreement for *mature*.

The qualitative analysis of the three videos led to the replacement of both the CS-upper and CS-lower, and a slight modification of the DS. The criterion was reviewed and the following statements were selected: CS-upper (forcefully stretch and reach with legs), and CS-lower (difficulty performing one-foot takeoff and land on opposite foot). In addition, the word "leg" was added to the DS. This was done based on input from the subjects.

In the CS-upper, the statement, *forceful extension of takeoff leg*, was replaced with *forceful stretch and reach with legs*. By watching the videos it was clear that, in some cases, the extension of the takeoff leg would occur even if there was no evidence of spring and elevation during the push-off. The CS-lower was also replaced. The original statement, *unable to push-off and gain distance & elevation*, was replaced with *difficulty performing one-foot takeoff and landing on opposite foot*. It is expected to be easier for observers to identify whether a performer is being inconsistent in taking off with one foot and landing with the opposite foot, than trying to interpret the subjectivity of the previously suggested statement. Thus, the two new statements are expected to better discriminate, and consequently improve accuracy of classifications, when the decision tree is used to classify elementary-level individuals.

Table 10: Final decision tree for leap

(DS) Intentional arm opposition **Changed to:** Intentional arm-leg opposition	True	(CS-upper) Forceful extension of takeoff leg **Changed to:** Forceful stretch and reach with legs	True	M
			False	E
	False	(CS-lower) Unable to push-off and gain distance & elevation **Changed to:** Difficulty performing one-foot takeoff and land on opposite foot	False	E
			True	I

Note. M= Mature, E= Elementary, I= Initial. Shaded areas denote the expected path. (DS) Discriminatory statement; (CS-upper) Upper confirmatory statement, (CS-lower) Lower confirmatory statement.

3.4 KICK

The proportion of specific agreement was $P_{s(1)}$= .28 for the initial-level, $P_{s(2)}$= .56 for the elementary-level, and $P_{s(3)}$= .91 for the mature-level categories. Presented next are the results for each of the two categories that had a P_s below .70.

The disagreement observed in the initial-level category was high in Video 2 (90%), but slightly lower in Video 6 (76.7%), and Video 7 (73.3%). There was a high percentage of disagreement in the left side of the decision tree for Video 2 (83.3%), but not for Video 6 (0%) or Video 7 (0%). Further, when reviewing all three videos, the mean percentage disagreement for the right side of decision tree was greater (52.2%) compared to the left side (27.8%). The disagreement in the left side of the decision tree was caused by Video 2 (83.3%), while the incorrect responses for Video 6 (76.7%) and Video 7 (73.3%) accounted for most of the disagreement on the right side of the decision tree.

The proportion of specific agreement for the elementary-level category of kick was also low (.56). When summarizing all three videos, the discrepancy was greater in the CS-lower (38.9%) compared to the CS-upper (6.7%). It should be noted that the percentage of respondents selecting the right-lower path of the decision tree was greater (72.2%) compared to those selecting the right-upper path (27.8%). However, the majority of those in Video 3 (13.3% out of 16.7%) and Video 4 (43.3% out of 53.3%) in the upper path correctly classified the videos as elementary. This suggests that the CS-upper worked as expected when used with elementary-level videos. The same cannot be said about the CS-lower.

Table 11: Summary of percentage agreement/disagreement for kick

	Videos			*Mean %*	
Category	*% Agreement for 1st Video*	*% Agreement for 2nd Video*	*% Agreement for 3rd Video*	*Agree-ment*	*Disa-greement*
Initial (2, 6, 7)					
DS					
True	83.3**	0.0**	0.0**	-	27.8_a
False	16.7*	100.0*	100.0*	72.2	-
CS-upper					
True	3.3**	0.0**	0.0**	-	1.1_b
False	76.7**	0.0**	0.0**	-	25.6_b
CS-lower					
True	10.0*	23.3*	26.7*	20.0	-
False	6.7**	76.7**	73.3**	-	52.2_c
Elementary (1, 3, 4)					
DS					
True	13.3	16.7	53.3	-	-
False	86.7	83.3	46.7	-	-
CS-upper					
True	6.7**	3.3**	10.0**	-	6.7_d
False	6.7*	13.3*	43.3*	21.1	-
CS-lower					
True	36.7**	66.7**	13.3**	-	38.9_e
False	50.0*	17.7*	33.3*	33.7	-
Mature (5, 8, 9)					
DS					
True	100.0*	100.0*	100.0*	100.0	-
False	0.0**	0.0**	0.0**	-	0.0_f
CS-upper					
True	80.0*	100.0*	90.0*	90	-
False	20.0**	0.0**	10.0**	-	10.0_g
CS-lower					
True	0.0**	0.0**	0.0**	-	0.0_h
False	0.0**	0.0**	0.0**	-	0.0_h

Note. DS= Discriminatory statement; CS= Confirmatory statement. Single asterisk denotes correct response. Double asterisk denote incorrect response. Numbers enclosed in parentheses denote the actual video numbers for each category. a= disagreement for initial-DS; b= disagreement for initial-CS-upper (adding up both values of *b* amounts the *a* value); c= disagreement for initial-CS-lower (adding up *a* and *c* amounts the total average disagreement for *initial*); d= disagreement for elementary-CS-upper; e= disagreement for elementary-CS-lower (adding up *d* and *e* amounts the total average disagreement for *elementary*); f= disagreement for mature-DS; g= disagreement for mature-CS-upper; h= disagreement for mature-CS-lower (adding up both values of *h* amounts the *f* value) – adding up *f* and *g* amounts the total average disagreement for *mature*.

The qualitative analysis of the six videos (three initial level and three elementary level) led to a major change in the decision tree for kick. First, the CS-lower criterion, *a pushing rather than a strike action is predominant*, was replaced with *no step is taken toward the ball*. In addition, the DS criterion switched places with the CS-upper (see Table 12). Further, the words "follow-through" were added to the beginning of the new DS. This was done based on input from the subjects during data collection.

Similar to the new CS-lower for leap, the CS-lower for kick is believed to be less subjective than the previously proposed statement. It may be easier for observers to differentiate between *stepping toward the ball* & *standing still behind the ball before kicking* than distinguish between a *pushing* & *striking pattern*. In addition to the replacement of the CS-lower, the DS and CS-upper switched places. It would be redundant to leave the statement *approach is either from a run or leap* in the left side of the decision tree, with the new CS-lower, *no step is taken toward the ball*. Therefore, the original DS was moved to the right-upper part of the decision tree and the original CS-upper was placed in the left side of the decision tree. This should not risk the ability of the left side of the decision tree to discriminate individuals between initial and mature levels because the former CS-upper, *support foot rises to toes or leaves the surface entirely*, is also believed to discriminate. It is unlikely that a performer who is at the initial level would be able to raise his/her foot to toes or lift it entirely from the surface. This is because for that to happen, one needs to approach the ball from a run or leap. And initial level performers are not likely to show such a pattern. They rely heavily on the extension of the knee to kick the ball, which is done in a standing position in front to the ball (Gallahue & Donnelly, 2003). In addition to the switch between DS and CS-upper, the words "follow-through" were added to the beginning of the statement. Therefore, the updated DS is: *follow-through w/ support foot rising to toes or leaving the surface*.

One last note is about the disagreement observed (27.8%) in the left side of the decision (DS), which was caused by Video 2. This could question the use of the DS in the updated version of the decision tree. In fact, when observing Video 2 closely, it was noticed the performer runs towards the ball, stops and then kicks it instead of running towards the ball and kicking without stopping, which was the expected pattern for a *true* response. The new DS criterion is believed to fix this problem.

Table 12: Final decision tree for kick

<table>
<tr><td rowspan="4">(DS)
Approach is either from a run or leap

Changed to:

Follow-through w/ support foot rising to toes or leaving surface</td><td rowspan="2">True</td><td rowspan="2">(CS-upper)
Support foot rises to toes or leaves the surface entirely

Changed to:

Approach is either from a run or leap</td><td>True</td><td>M</td></tr>
<tr><td>False</td><td rowspan="2">E</td></tr>
<tr><td rowspan="2">False</td><td rowspan="2">(CS-lower)
A pushing rather than a striking action is predominant

Changed to:

No step is taken toward the ball</td><td>False</td></tr>
<tr><td>True</td><td>I</td></tr>
</table>

Note. M= Mature, E= Elementary, I= Initial. Shaded areas denote the expected path. (DS) Discriminatory statement; (CS-upper) Upper confirmatory statement, (CS-lower) Lower confirmatory statement.

3.5 HAND DRIBBLE

The proportion of specific agreement was $P_{s(2)}$= .65 for the elementary-level, $P_{s(1)}$= .78 for the initial-level, and $P_{s(3)}$= .84 for the mature-level categories. Thus, further steps were taken to identify the source of the disagreements for the elementary-level category.

The percentage agreement was greater for Video 2 (90.0%) when compared to Video 4 (56.7%), and Video 6 (36.6%). Regarding the location of the disagreement, Table 13 shows that the disagreement in the CS-upper (30.0%) was greater than the disagreement observed in the CS-lower (7.8%).

Table 13: Summary of percentage agreement/disagreement for hand dribble

	Videos			*Mean %*	
Category	*% Agreement for 1st Video*	*% Agreement for 2nd Video*	*% Agreement for 3rd Video*	*Agree-ment*	*Disa-greement*
Initial (3, 7, 8)					
DS					
True	16.7**	30.0**	23.3**	-	23.3_a
False	83.3*	70.0*	76.7*	76.7	-
CS-upper					
True	0.0**	0.0**	0.0**	-	0.0_b
False	16.7**	30.0**	23.3**	-	23.3_b
CS-lower					
True	76.7*	53.3*	73.3*	67.8	-
False	6.7**	16.7**	3.3**	-	8.9_c
Elementary (2, 4, 6)					
DS					
True	3.3	70.0	80.0	-	-
False	96.7	30.0	20.0	-	-
CS-upper					
True	0.0**	43.3**	46.7**	-	30.0_d
False	3.3*	26.7*	33.3*	21.1	-
CS-lower					
True	6.7**	0.0**	16.7**	-	7.8_e
False	90.0*	30.0*	3.3*	41.1	-
Mature (1, 5, 9)					
DS					
True	100.0*	96.7*	93.3*	96.7	-
False	0.0**	3.3**	6.7**	-	3.3_f
CS-upper					
True	96.7*	96.7*	90.0*	94.5	-
False	3.3**	0.0**	3.3**	-	2.2_g
CS-lower					
True	0.0**	0.0**	0.0**	-	0.0_h
False	0.0**	3.3**	6.7**	-	3.3_h

Note. DS= Discriminatory statement; CS= Confirmatory statement. Single asterisk denotes correct response. Double asterisk denote incorrect response. Numbers enclosed in parentheses denote the actual video numbers for each category. a= disagreement for initial-DS; b= disagreement for initial-CS-upper (adding up both values of *b* amounts the *a* value); c= disagreement for initial-CS-lower (adding up *a* and *c* amounts the total average disagreement for *initial*); d= disagreement for elementary-CS-upper; e= disagreement for elementary-CS-lower (adding up *d* and *e* amounts the total average disagreement for *elementary*); f= disagreement for mature-DS; g= disagreement for mature-CS-upper; h= disagreement for mature-CS-lower (adding up both values of *h* amounts the *f* value) – adding up *f* and *g* amounts the total average disagreement for *mature*.

The qualitative analysis of all three elementary-level videos led to the replacement of the CS-upper criterion, *controlled directional dribble*, with *visual monitoring unnecessary*. Table 14 shows the updated decision tree for hand dribble.

The main reason for the replacement of the CS-upper statement was because the original statement demanded a great deal of interpretation as with the new statement the observer simply needs to decide whether or not the performer relies on visual monitoring during the action. This should help to improve the accuracy of classifications of the hand dribble decision tree whenever it is used to classify elementary-level individual.

Table 14: Final decision tree for hand dribble

<table>
<tr><td rowspan="4">(DS)
Ball is pushed (not struck) toward ground</td><td rowspan="2">True</td><td rowspan="2">(CS-upper)
Controlled directional dribble

Changed to:

Visual monitoring unnecessary</td><td>True</td><td>M</td></tr>
<tr><td>False</td><td rowspan="2">E</td></tr>
<tr><td rowspan="2">False</td><td rowspan="2">(CS-lower)
Repeated bounce and catch pattern</td><td>False</td></tr>
<tr><td>True</td><td>I</td></tr>
</table>

Note. M= Mature, E= Elementary, I= Initial. Shaded areas denote the expected path. (DS) Discriminatory statement; (CS-upper) Upper confirmatory statement, (CS-lower) Lower confirmatory statement.

3.6 OVERHAND THROW

Steps were taken to identify the source of the disagreements for the elementary-level category, since its index of proportion of specific agreement was below .70 ($P_{s(2)}$= .50).

The percentage agreement was greater for Video 9 (86.7%) when compared to Video 1 (10%), and Video 2 (10%). Regarding the location of the disagreement, Table 15 shows that the mean percentage disagreement in the CS-lower (63.3%) was greater than the disagreement observed in the CS-upper (1.1%).

Table 15: Summary of percentage agreement/disagreement for overhand throw

	Videos			*Mean %*	
Category	*% Agreement for 1st Video*	*% Agreement for 2nd Video*	*% Agreement for 3rd Video*	*Agree-ment*	*Disa-greement*
Initial (4, 5, 6)					
DS					
True	0.0**	0.0**	0.0**	-	0.0_a
False	100.0*	100.0*	100.0*	100.0	-
CS-upper					
True	0.0**	0.0**	0.0**	-	0.0_b
False	0.0**	0.0**	0.0**	-	0.0_b
CS-lower					
True	96.7*	100.0*	93.3*	96.7	-
False	3.3**	0.0**	3.3**	-	2.2_c
Elementary (1, 2, 9)					
DS					
True	3.3	3.3	80.0	-	-
False	96.7	96.7	20.0	-	-
CS-upper					
True	0.0**	0.0**	3.3**	-	1.1_d
False	3.3*	3.3*	76.7*	27.8	-
CS-lower					
True	90.0**	90.0**	10.0**	-	63.3_e
False	6.7*	6.7*	10.0*	7.8	-
Mature (3, 7, 8)					
DS					
True	100.0*	100.0*	93.3*	97.8	-
False	0.0**	0.0**	6.7**	-	2.2_f
CS-upper					
True	96.7*	100.0*	93.3*	96.7	-
False	3.3**	0.0**	6.7**	-	3.3_g
CS-lower					
True	0.0**	0.0**	6.7**	-	2.2_h
False	0.0**	0.0**	0.0**	-	0.0_h

Note. DS= Discriminatory statement; CS= Confirmatory statement. Single asterisk denotes correct response. Double asterisk denote incorrect response. Numbers enclosed in parentheses denote the actual video numbers for each category. a= disagreement for initial-DS; b= disagreement for initial-CS-upper (adding up both values of *b* amounts the *a* value); c= disagreement for initial-CS-lower (adding up *a* and *c* amounts the total average disagreement for *initial*); d= disagreement for elementary-CS-upper; e= disagreement for elementary-CS-lower (adding up *d* and *e* amounts the total average disagreement for *elementary*); f= disagreement for mature-DS; g= disagreement for mature-CS-upper; h= disagreement for mature-CS-lower (adding up both values of *h* amounts the *f* value) – adding up *f* and *g* amounts the total average disagreement for *mature*.

The qualitative analysis of all three elementary-level videos led to the replacement of the CS-lower criterion, *purposeless shift of feet or remain stationary*, with *action is mainly from elbow and resembles a push*. Table 16 shows the updated decision tree for overhand throw.

The replacement was mainly due to the inability of the original statement, *purposeless shift of feet or remain stationary,* to discriminate between initial and elementary-level individuals. This possibly has to do with the high subjectivity of the term "purposeless shift of feet" in the first part of the CS-lower. The new statement (action is mainly from elbow and resembles a push) may be easier to be identified when compared to the previously proposed statement.

Table 16: Final decision tree for overhand throw

<table>
<tr><td rowspan="4">(DS)
As weight is shifted, there is a step w/ opposite foot</td><td rowspan="2">True</td><td rowspan="2">(CS-upper)
Trunk markedly rotates to throwing side during preparation action</td><td>True</td><td>M</td></tr>
<tr><td>False</td><td rowspan="2">E</td></tr>
<tr><td rowspan="2">False</td><td rowspan="2">(CS-lower)
Purposeless shift of feet or remain stationary

Changed to:

Action is mainly from elbow and resembles a push</td><td>False</td></tr>
<tr><td>True</td><td>I</td></tr>
</table>

Note. M= Mature, E= Elementary, I= Initial. Shaded areas denote the expected path. (DS) Discriminatory statement; (CS-upper) Upper confirmatory statement, (CS-lower) Lower confirmatory statement.

3.7 OTHER CHANGES

In addition to the changes made to the decision trees that were selected based on the P_s values of the respective categories, there was a change to the decision tree for skip. The P_s values for all the categories of skip were above .70 ($P_{s(1)}$= .93, $P_{s(2)}$= .77, $P_{s(3)}$= .74). However, a change was made to the CS-lower of the decision tree (see Table 17). Subjects often asked the meaning of the word “arhythmically” during data collection, which is part of the statement. Therefore, it was changed to “off-beat”. Further, by comparing the videos with the decision tree, the word “and” in the middle of the statement may, in certain occasions, be misleading. An elementary-level performer may be off-beat, but not necessarily fast pace. Thus, the word “AND” was substituted with the word “OR”.

Table 17: Final decision tree for skip

(DS) Arms move rhythmically in opposition to legs	True	(CS-upper) Low vertical lift on hop	True	M
			False	E
	False	(CS-lower) Arrhythmical and fast pace **Changed to:** Off-beat OR fast pace	False	
			True	I

Note. M= Mature, E= Elementary, I= Initial. Shaded areas denote the expected path. (DS) Discriminatory statement; (CS-upper) Upper confirmatory statement, (CS-lower) Lower confirmatory statement.

4.0 DISCUSSION

Painter (1994) has called for research to determine how to modify the developmental movement sequences into more functional assessment instruments for practitioners. In addition, the range of movement behaviors should be limited to facilitate the observational process (Painter, 1994). This study was conducted to do just that: to develop a rating system containing only a few, yet key, performance criteria and gather information to determine whether teachers could use this scale (composite decision tree) to evaluate the skill levels of children on six object-control and five locomotor fundamental movement skills.

The major challenge in the development of such a rating system is to select the performance criteria that works as strong discriminators thus allowing for accurate classification of individuals into their membership groups (e.g., initial, elementary, mature). Taylor (1979) warned about the importance of selecting only those features from the developmental sequences that are crucial to movement efficiency. Thus, the changes that were made to the original proposed rating system (composite decision trees) reflected Taylor's suggestion.

Overall, seven decision trees (side slide, horizontal jump, leap, kick, hand dribble, overhand throw, and skip) underwent modifications, while four (hop, catch, strike, and batting) remained unchanged from the original format. The number of changes varied for each tree.

The changes performed to the decision trees is divided into two categories: (1) changes due to problems with the format of the statements (structural changes), and (2) changes due to lack of objectivity of the statement. Structural changes include adding, deleting, and/or substituting words within the statement. The second type of modification yielded a complete replacement of the statement with the goal of making it a better discriminator between any two levels of performance. The structural changes are discussed next.

The decision trees for side slide and skip underwent structural change modifications. For side slide, the CS-upper was changed because the disagreement was caused by the double negative in the statement. In fact, in eight occasions, subjects indicated that the double negative was confusing when responding true/false to that statement. Similarly, the CS-lower for skip was changed based on the fact that 11 subjects asked for clarification with regards to the word "arrhythmical" during the training session. Changes due to a possible lack of objectivity of the statement are discussed next.

The main reason that led to the modifications of the other five decision trees (horizontal jump, leap, kick, overhand throwing, and hand dribble) was due to the subjectivity of the statements. Painter (1994) addressed the issue of subjectivity stating that the distinct behaviors of a given developmental sequence should be differentiated from the less observable behaviors when selecting a performance criterion for a skill. The lack of objectivity of a statement will negatively affect the power to work as a discriminator within the decision tree. For example, the mean percentage of disagreement at the CS-lower, *trunk moves in vertical direction; little emphasis on length of jump*, for the horizontal jump was 36.7%. This indicates that observers were rating certain videos differently. One plausible explanation is that the statement failed to work as a discriminator between the elementary and initial levels due to the lack of objectivity. This is a possible problem associated with the decision tree itself.

Another explanation for the amount of variance associated with ratings of horizontal jump is the number of trials. It is possible that subjects needed more than three trials to detect differences in skill performance. Ulrich and Branta (1988) employed a generalizability analysis to investigate the minimal number of trials needed to obtain reliable developmental stages for hop, horizontal jump, and run using the Michigan State University developmental sequences with developmental stages from 1 through 4. The generalizability analysis was followed by nine decision studies (D-studies) for each skill, using nine different sets of conditions (i.e., one observer and one trial; two observers and three trials; and so on). The authors established a coefficient generalizability of .80 as the minimum condition of observation. Their results indicated that hop was the only skill than can be reliably rated in a teaching situation where one observer is present. A generalizability coefficient of .88 was achieved when ratings were averaged over one observer and three trials. The jump required three observers and three trials while the run required three observers and more than five trials. The authors concluded that, presumably, more trials would reach the desired criterion with **one** observer for the horizontal jump. Exactly how many trials are necessary for horizontal jump to be reliably rated in a teaching situation is unclear (Ulrich & Branta, 1988).

It is difficult to extend the discussion regarding the number of trials to the other five skills (side slide, leap, kick, hand dribble, and overhand throw) selected for further analysis. This is due to the lack of studies that seek to identify the minimum number of trials necessary for reliable ratings of individuals in typical teaching situations for the skills above. In the lack of such studies, it could be speculated that subjects in the current study needed more than three trials to detect differences in performance, at least for leap, kick, and overhand throw. This is because unlike side slide and hand dribble, the trials for leap, kick and overhand throw are brief and might require more training for differences to be detected. In hand dribble, performers were asked to dribble the ball for 20 seconds while standing. In side slide, performers were given three trials to travel a distance of 12 meters for each trial. Although possible, this claim might not itself explain the low agreement on the skills of leap, kick, and overhand throw. This is because agreement was high on other three skills (catch, batting, strike) in which the trials are also brief.

One final plausible explanation for the amount of variation observed in the six skills identified for further analysis has to do with the misconception that the performance of motor skills is age-dependent. This was also considered in the Ulrich and Branta's (1988) study to explain their results. Although subjects in the current study were warned about making such mistakes, it could be that observers were more prone to give higher ratings (mature) to older looking children and a lower rating (initial) to younger children despite the body actions being displayed. Ulrich and Branta (1988) suggested that this might be reduced through training. Arguably, observers could lose this misconception if more examples of older children performing at the lower level and younger children performing at the higher level are used during training.

One last issue that needs to be addressed is the fact that the elementary-level category was most often selected for further analysis. The initial-level category was the source of disagreement in only two skills, and the mature-level category was not identified as being the source of disagreement. However, the elementary-level category was the major source of disagreement in six out of the eleven skills proposed for the FG-COMPASS.

The fact that the elementary-level category was most often selected for further analysis was not a surprise. Mills (1983) discussed the difficulty involved in classifying the "borderline performers". Individuals do not fall under discrete categories with regards to their fundamental movement development. Discrete categories are used in order to simplify the process of assessment. Consider the concept of a continuum regarding fundamental movement skill development. On the left side of this continuum there would be the initial-level category discussed in this study. In the middle, there would be the elementary-level; and the mature-level would be placed on the right. Now consider that minus signs are placed to the left of the initial-level category and plus signs place at the right. Thus, Johnny might be at the initial low level of skip (i.e., sequence of minus signs to the left of the initial-level category). Now consider Sara being more advanced, but not advanced enough to be considered at the elementary level. She would be placed in between the initial and the elementary-level categories (i.e., sequence of plus signs to the right of the initial level). Now, the same concept can be used for the mature level with the sequence of minus signs representing the least mature level and the sequence of plus signs as the most advanced mature level. Using the same concept, the elementary level can also be considered more and less advanced. Thus, there is a greater chance of an elementary-level individual being misclassified if he/she is either more towards the left of right of the in the continuum, compared to either an initial or mature-level individual. An initial-level individual can only, at least in theory, be misclassified if he/she is more towards the right of the continuum (sequence of plus signs). The same applies for the mature level. The problem just discussed can be exacerbated if either the CS-upper or CS-lower are not functioning as expected within the decision tree.

In summary, the decision trees for seven skills (side slide, horizontal jump, leap, kick, hand dribble, overhand throw, and skip) underwent modifications as a result of this study. The other four decision trees (catch, strike, batting, and hop) remained unchanged. The changes were justified mainly on the basis of structural problems or lack of objectivity of the statements. An additional explanation for the amount of variance in ratings refers to the number of trials necessary for observers to detect differences in performance. This is plausible at least for the skills in which the trials are short lasting in nature, which is the case of horizontal jump, leap, and kick. Additionally, the possibility of age-dependent misconceptions was discussed. It could be that observers in the current study were inclined to give higher ratings to older looking children and lower ratings to younger looking children. Finally, a rationale was provided to explain why the elementary-level category was most often selected for further analysis.

4.1 LIMITATIONS AND FUTURE RESEARCH

An inherent limitation of studies in which observers make judgments on the basis of skill performance is the high level of subjectivity in the ratings. Even though training was provided prior to the testing session, it is difficult to ensure subjects are using the same trait when rating the videos. An additional limitation of the present study is the fact that each subject had to rate 110 videos. Although subjects were given two 5-minute breaks during the testing session, it is possible that fatigue influenced the results. One alternative would be to recruit more than 30 subjects, so that each subject had to rate fewer videos.

The outcome of the current study provides initial evidence that the decision trees (rating scale) developed for each of the eleven skills can be used to classify individuals into their membership groups with regard to fundamental movement skill performance. The next step is to verify whether the changes made to the six skills (side slide, horizontal jump, leap, kick, hand dribble, overhand throw), which were selected for further analysis, will help to improve the accuracy of classifications. Therefore, a follow-up study needs to be conducted which will involve only these six skills.

Then, assuming that no more changes are necessary to the six skills above, the next logical approach would be to determine whether this newly developed rating scale can reliably be used by practitioners in the field. To do that, reliability and objectivity studies need to be conducted. Reliability studies can be carried out to evaluate the dependability of the scores yielded by administering the FG-COMPASS. The stability (test-retest) reliability procedure can be used. In addition, objectivity (rater reliability) studies can be performed in which the degree of agreement between raters is evaluated.

APPENDIX A

FINAL DECISION TREES

The following are the updated decision trees after this study was conducted.

Decision trees for Locomotor

HOP	Swing leg moves like a pendulum	T	Balance is well controlled	T	M
				F	E
		F	Swing leg held in front of body	F	E
				T	I
HORIZONTAL J.	Arms thrust forward forcefully on takeoff	T	Arms move high and to rear during preparatory crouch	T	M
				F	E
		F	Difficult in using both feet	F	E
				T	I
SKIP	Arms move rhythmically in opposition to legs	T	Low vertical lift on hop	T	M
				F	E
		F	Off-beat OR fast pace	F	E
				T	I
SIDE SLIDE	Smooth, rhythmical action	T	Clear airborne throughout	T	M
				F	E
		F	Double hop or step occurs	F	E
				T	I
LEAP	Intentional arm-leg opposition	T	Forceful stretch and reach w/ legs	T	M
				F	E
		F	Difficulty performing one-foot takeoff and land on opposite foot	F	E
				T	I

Key: T= True; F= False; I= Initial; E= Elementary; M=Mature

Decision trees for Object-Control

OVERHAND THROW	As weight is shifted, there is a step with opposite foot	T	Trunk markedly rotates to throwing side during preparation action	T	M
				F	E
		F	Action is mainly from elbow and resembles a push	F	E
				T	I
KICK	Follow-through w/ support foot raising to toes or leaving surface	T	Approach is either from a run or leap	T	M
				F	E
		F	No step is taken toward the ball	F	E
				T	I
HAND DRIBBLE	Ball is pushed (not struck) toward ground	T	Visual monitoring unnecessary	T	M
				F	E
		F	Repeated bounce and catch pattern	F	E
				T	I
CATCH	Catching resembles a scooping action	F	Well-timed and simultaneous motion in hands grasp	T	M
				F	E
		T	Hands not utilized in catching action	F	E
				T	I
STRIKING	The swing arm moves through a full range of motion	T	Steps into the swing w/ a differentiated trunk-hip rotation	T	M
				F	E
		F	Swing is down (vertical plane) rather than sideways	F	E
				T	I
BATTING	Striking occurs in a long, full arc in a horizontal plane	T	Differentiated trunk-hip rotation	T	M
				F	E
		F	Motion is from back to front in a downward plane	F	E
				T	I

APPENDIX B

INFORM CONSENT FOR VIDEOTAPING

Parent/Guardian PHOTOGRAPHIC/VIDEOTAPE CONSENT FORM

Procedure:
Your son/daughter will perform 13 tasks commonly use in Physical education instruction. The tasks are listed below:

Locomotor Skills

- Leap
- Slide
- Skip
- Horizontal jump
- Hop

Object-control Skills

- Catch
- Side-Arm Strike
- Batting
- Hand dribble
- Kick
- Overhand Throw

The tasks above will be administered by my child's physical education (PE) teachers at Falk School during regular PE classes.
I, ______________________________________, give permission to the University of Pittsburgh to take and use photographs/videos of my child ____ performing fundamental movement skills for educational and research purposes including, but not limited to, use in University classes and research projects.

I understand that I will not be paid for these photographs/videos and have no right to them. I release the University of Pittsburgh, its employees, and its agents from any and all claims whatsoever of harm or otherwise that may occur from showing, using, or distributing these photographs.

I have read this form or have had it read to me. I understand that it says and agree to its terms.

Signed: ____
Date:____

Parent or Guardian (if under 18): ____

Witness: _______________________________________
Date: ____

APPENDIX C

VIDEO/SKILL RANDOMIZATION

S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	S 9	S 10	S 11	S 12	S 13	S 14	S 15	S 16	S 17	S 18	S 19	S 20	S 21	S 22	S 23	S 24	S 25	S 26	S 27	S 28	S 29	S 30
11	6	1	5	6	7	2	2	5	7	4	1	6	2	10	5	9	3	1	10	10	9	11	6	10	6	9	5	2	8
8	1	10	2	5	3	3	10	9	8	5	10	11	11	8	8	5	11	2	1	7	4	6	3	8	3	6	3	11	5
6	8	7	10	8	8	11	4	3	11	10	7	10	3	2	3	4	1	9	4	4	7	5	11	11	9	8	8	6	1
1	11	4	1	11	4	9	5	6	3	7	3	1	6	11	2	8	2	5	11	3	6	3	8	1	1	5	2	10	11
4	3	8	8	4	9	8	9	4	5	1	2	5	1	4	7	11	5	7	2	1	10	10	1	4	7	3	9	7	2
7	4	6	6	3	1	1	7	8	10	9	6	7	5	5	11	1	10	10	9	5	8	7	4	7	5	2	11	1	7
2	10	2	9	7	5	6	1	7	9	3	5	9	9	7	10	6	4	8	6	8	2	4	2	6	10	7	10	9	10
3	5	11	7	2	6	7	3	1	4	2	9	4	8	3	9	2	9	4	7	11	5	1	9	3	4	10	7	4	9
5	7	9	4	9	2	5	6	11	1	8	8	8	7	9	6	10	8	11	3	6	1	8	7	5	8	4	6	8	4
10	2	3	11	10	11	10	11	2	2	11	4	2	10	1	1	3	7	3	8	2	3	9	10	2	11	1	4	3	6
9	9	5	3	1	10	4	8	10	6	6	11	3	4	6	4	7	6	6	5	9	11	2	5	9	2	11	1	5	3
1	4	1	7	3	1	4	2	3	2	1	8	7	8	7	8	9	1	9	5	1	5	6	2	7	5	5	4	3	5
4	8	5	3	9	5	9	3	5	4	7	5	9	4	3	9	2	5	8	3	3	2	7	5	8	3	4	7	1	6
7	9	7	5	6	8	2	5	7	6	2	7	2	6	9	7	1	3	1	9	4	1	4	7	6	6	7	1	7	8
3	7	2	2	7	6	1	1	2	1	6	6	1	5	4	3	4	8	4	4	7	4	3	9	4	4	3	9	8	2
9	2	6	8	2	2	3	9	1	8	5	1	8	3	1	2	7	7	5	1	5	6	9	6	5	2	1	6	6	1
5	1	3	9	4	3	7	7	9	3	4	4	4	9	5	4	8	2	2	8	8	8	8	8	9	1	9	3	9	4
8	6	8	6	1	4	6	8	8	7	3	3	6	1	2	1	5	9	3	2	6	3	1	4	1	8	2	2	5	3
2	5	4	1	5	9	8	6	4	5	8	9	5	7	8	6	6	6	7	6	9	9	2	1	3	9	6	8	4	7
6	3	9	4	8	7	5	4	6	9	9	2	3	2	6	5	3	4	6	7	2	7	5	3	2	7	8	5	2	9

1= Hop; 2= Side Slide; 3= Horizontal jump; 4= Leap; 5= Skip; 6= Catch; 7= Kick; 8= Strike; 9= Hand dribble; 10= Batting; 11= Overhand Throw

APPENDIX D

AGE DISTRIBUTION ACROSS SKILL AND SKILL LEVEL

SKILL	LEVEL	AGE (Months)			M(L)	SD(L)	M(S)	SD(S)
Hop	1	80	74	120	91.33	25.01	103.00	24.99
	2	111	140	72	107.67	34.12		
	3	105	93	132	110.00	19.97		
Horizontal jump	1	74	135	66	91.67	37.74	113.67	30.23
	2	141	149	89	126.33	32.58		
	3	124	131	114	123.00	8.54		
Skip	1	74	78	134	95.33	33.55	100.78	25.18
	2	73	81	114	89.33	21.73		
	3	130	99	124	117.67	16.44		
Side Slide	1	82	72	74	76.00	5.29	106.00	27.85
	2	97	140	135	124.00	23.52		
	3	93	130	131	118.00	21.66		
Leap	1	118	83	91	97.33	18.34	117.00	19.47
	2	124	108	140	124.00	16.00		
	3	124	130	135	129.67	5.51		
Overhand Throw	1	73	81	100	84.67	13.87	97.78	15.26
	2	96	120	105	107.00	12.12		
	3	100	116	89	101.67	13.58		
Catch	1	74	86		80.00	8.49	85.25	15.21
	2	83	66	73	74.00	8.54		
	3	101	112	87	100.00	12.53		
Hand Dribble	1	73	66	107	82.00	21.93	101.22	22.22
	2	81	123	121	108.33	23.69		
	3	112	123	105	113.33	9.07		
Bating	1	94	91	97	94.00	3.00	116.22	19.72
	2	102	135	128	121.67	17.39		
	3	130	129	140	133.00	6.08		
Striking	1	91	94		92.50	2.12	109.25	20.14
	2	99	81	120	100.00	19.52		
	3	128	131	130	129.67	1.53		
Kick	1	83	70	78	77.00	6.56	98.22	27.09
	2	73	77	135	95.00	34.70		
	3	124	132	112	122.67	10.07		

Note: M(L)= Mean for each level; SD(L)= Standard deviation for each level; M(S)= Mean for each skill; SD(S)= Standard deviation for each skill.

APPENDIX E

TELEPHONE TRANSCRIPT

"Hello [potential participant's name], my name is Ovande Furtado and I am the principal investigator for the research study you contacted me about.

- I am calling in response to your [email] [phone call].
- I am calling because you [left a message] [sent an email] showing interest in participating as subject in the research study title " Development and Validation of the FG-COMPASS.

QUESTION 1
- Do you still interested in participating as subject in this study?

  [IF NO]
  Thank you. Good-bye.
  [IF YES]
  Great! Before I begin giving you more information about this study, I have to ascertain that you qualify to participate as subject. Thus, I will ask you a few questions [ASK QUESTION 2].

QUESTION 2
- Are you currently enrolled in or graduated from a K-12 Physical Education teacher certification program?

  [IF YES]
  Great! You pre-qualify to participate as subject in this research study. I will now ask you a few other questions to ascertain your eligibility. [Skip to QUESTION 3]
  [IF NO]
  Unfortunately, you do not qualify to participate as subject in this research. Only individuals who are currently enrolled or have graduated from either an undergraduate or graduate physical education teacher certification program are entitled to participate in this research study. Good-bye.

QUESTION 3

- Participation in this study involves classifying 110 video clips that depicts students performing 11 fundamental movement skills (e.g., galloping, catch, etc.). Your task will be to watch each video-clip and classify them according to three developmental levels; namely, initial level, elementary level and mature level. Your ratings for each video-clip will be compared to the actual developmental level of the children appearing on the video, which was previously determined. Participation in this study will take approximately 4 hours of your time. First you will come for a training session that is expected to last 2 hours. Then five days later, you will come for the testing session, which is also expected to last approximately 2 hours. Both the training and testing sessions will be carried-out at [redacted].
In appreciation of your time commitment, you will receive $40 upon the completion of the study. In addition, you can earn an extra $50 based on your performance rating the video clips. The two best-ranked participants will receive the extra $50.

Now, based on what I just said, do you feel that you are able to participate as subject in this research study?

[IF NO]
"Thank you, good-bye".
[IF YES]
"Thank you; we appreciate your interest in our research [PROCEED].

- I have a session open on [day and date] at [time, a.m. or p.m.]. Will you be available then? You will need to come in about 15 minutes early".

[IF NO]
Offer another day and time until one is found that is mutually convenient.
[IF YES]
"This is great. Let me give you some important details about the study [PROCEED].

- Have you got a pen so that you can write this down and keep it with you?"

"You should go to the 2nd floor of the Lantz building at [15 minutes before the time scheduled] on [mention day and date again]. Look for the [redacted] main office (suite 2506). Let the secretary know that you are there for the research study conducted by Professor Furtado. I will meet you inside the office.

Now, let me write down your email address. I will be sending this information via email just in case.

I will be sending this info via email just in case. The day before your session, I will contact you by email as a reminder. However, in the meantime, if you discover you will be unable to make it, please call me at 217-[redacted] and leave a message if I am not available or email me at [redacted]. Please try to provide at least 24 hours notice so that I can book another participant into that time slot and avoid losing lab time.

"I look forward to meeting you on [mention day, date and time again]. Thank you very much again for helping us with our research".

APPENDIX F

TRAINING TRANSCRIPT

Introduction

I will briefly demonstrate how to complete the training session. You may stop me at any time during this explanation to ask questions. Before participating as subject in this study; you will be required to complete a training session, consisting of 11 modules; one module for each one of the eleven skills that are tested in this study.

General Information

The purpose of the training session is to prepare you for the testing session that will take place 3 to 5 days from this day. You will watch video clips of children, of different skill levels, performing several fundamental movement skills, including: hop, skip, catch, kick, etc.

Your task during this training session is to:

First, watch each video. Second, answer “true” or “false” to two statements that will be made about the videos.

Visual cues

You will be given visual cues to which you are encouraged to focus your attention while watching the videos. For example: for the skill of hop, you should focus on the swing leg, and the overall balance during the performance. This will help you when answering true or false to each statement.

Important

Each child on the videos will perform the same task 3 times. Your goal is to look for consistency of performance.

Demonstration

Next, I will demonstrate how to complete a training module. For the purpose of this demonstration we will complete a module using the skill of catch. So let's go ahead and start!

Using the computer in front of you, tap the link titled DEMO. This will take you to a page containing the visual cues for the skill you are currently rating. In the middle of the page you will see two or three visual cues that you should focus your attention while watching the videos for this skill. You should try to divide your attention among those visual cues while watching the three trials for each video, instead of looking at one specific visual cue in each trial.

Once you are done studying the visual cues, click on the button titled: CLICK HERE TO BEGIN.

The screen that you see now has two main areas: The top area and the bottom area. The top area is where the video clips will be played. The bottom area displays information regarding the statements.

Now, here is what you would do to complete the demo module:

First, you should read the first statement, which is displayed on the bottom of the screen.

Second, you should play the video-clip. After watching the video-clip, answer "true" or "false" for statement # 1. You will only be allowed to watch the video once.

Now say your answer for statement # 1 was in accordance with the criterion. If so, you will be presented with statement #2, and you should proceed in the same way you did for statement #1.

If, on the other hand, your answer was different from that of the criterion, then a red box appears next to the statement asking you to watch the video again and reconsider your answer. Ask me for clarification if, after watching the video for the second time, you still disagreeing with the criterion. Otherwise, change your response and move on to the next screen. Once you change your response, in this case from true to false, the system allows you to proceed answering the next statement. This will continue until you are done rating all three video clips. When you finished rating video-clip #3, you will be presented with message saying successfully completed the training for that given skill. You will be directed to the "Home Screen" again where you will select another skill.

Should you have any question during this training, do not hesitate to ask me for clarification.

APPENDIX G

TESTING TRANSCRIPT

This testing session is similar to the training session. The differences are: (1) videos will be displayed on the big screen, not on your computer; (2) you will use a pen to tap the computer screen and answer “true” or “false” to each statement presented; and (3) no feedback will be provided after you answer each statement.

During this testing session you will watch videos of children, of different skill levels, performing several fundamental movement skill tasks, including: hop, skip, catch, kick, etc.

Your task during this training session is to:

First: watch each video;
Second: answer, TRUE, OR FALSE to, 2 statements that will be made about the videos.

You will be given visual cues to which you have to focus your attention, while watching the videos. For example: for the skill of hop, you should focus on the swing leg, and the overall balance during the performance. This will help you when answering true or false to each statement.

Each child on the videos will perform the same task 3 times. Your goal is to look for consistency of performance.

Here is an example:

From the introduction's page, you will click on DEMO. This will take you to the VISUAL CUES PAGE.

In the middle of the page you will see 2 to 3 visual cues that you should focus your attention while watching the videos for a given skill.

Once you are done, click on "CLICK HERE TO BEGIN".

The screen that you see now has 3 main areas: The top area, the middle area, and the bottom area.

The top area is where the information regarding the module, video, and statement are displayed. The middle area is where the video will be played. The bottom area displays information regarding the statements.

Now, here is what you would do complete the demo module:

Now let’s focus on the bottom area. A single rectangle in the center means the statement is being presented for the first time.

The upper half of the rectangle will display either statement 1 or statement 2. Then, right below it the statement itself.

Based on your response to the statement, a new screen will be presented. For example:

Say, after watching video 1, your answer is TRUE for statement 1.

A message appears on the left hand-side of the bottom area.

The message will instruct you to watch the video for the second time, and reconsider your response.

Ask the principal investigator for clarification if after watching the video for the second time, you still disagreeing with the given response.

Otherwise, change your response and move on to the next screen.

Once you change your response, in this case from "true" to "false", the next screen is presented.

This is simply a confirmation that your response is in accordance with the skill level of the child depicted on the video.

By clicking on "statement 2", the screen with the second statement for video 1 is presented. It is similar to the first screen presented for statement 1. The difference is that now you are being asked to respond to a different statement.

When you are done watching all 3 videos for a module, this screen will be presented. It will instruct you to go back to the "HOME" page. There, you will be able to select another module.

Should you have any question during this training, do not hesitate to ask me for clarification.

APPENDIX H

KEY WORDS FOR FOCUS

HOP
- Swing leg
- Overall balance

SIDE SLIDE
- Smoothness
- Pace of action
- Height of flight

HORIZONTAL JUMP
- Arms
- Jumping distance

LEAP
- Arms
- Take-off leg

SKIP
- Arms
- Vertical lift

CATCH
- Arms
- Hand grasp

KICK
- Approach
- Support foot
- Striking leg

STRIKE
- Arms
- Trunk
- Direction of swing

HAND DRIBBLE
- Dribbling hand
- Control of dribble

BATTING
- Plane of striking
- Feet

OVERHAND THROW
- Feet
- Trunk rotation

BIBLIOGRAPHY

AERA. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Baquet, G., Twisk, W., Kemper, C., Van Praagh, E., & Berthoin, S. (2006). Longitudinal follow-up of fitness during childhood: interaction with physical activity. *American Journal of Human Biology*, *18*(1), 51-58.

Barrow, H. M., McGee, R., Tritschler, K. A., & Barrow, H. M. (1989). *Practical measurement in physical education and sport*. Philadelphia: Lea & Febiger.

Baumgartner, T. A., Mahar, M. T., Jackson, A. S., & Rowe, D. A. (2007). *Measurement for Evaluation in Physical Education and Exercise Science*. New York: McGraw-Hill.

Branta, C., Haubenstricker, J., & Seefeldt, V. (1984). Age changes in motor skills during childhood and adolescence. *Exercise and sport sciences reviews*, *12*, 467-520.

Burton, A. W., & Miller, D. E. (1998). *Movement Skill Assessment*. Champaign, IL: Human Kinetics.

Clark, J. (2007). On the problem of motor skill development. *Journal of Physical Education and Recreation*, *78*(5), 39-43.

Clark, J. E., & Phillips, S. J. (1985). A developmental sequence of the standing long jump. In J. E. Clark & J. H. Humphrey (Eds.), *Motor development : current selected research, vol.1*. Princeton, N.J.: Princeton Book Co., Pub.

Foley, J. T., Harvey, S., Chun, H., & Kim, S. (2008). The relationships among fundamental motor skills, health-related physical fitness, and body fatness in South Korean adolescents with mental retardation. *Research Quarterly for Exercise and Sport*, *72*(2), 149-57.

Furtado, Jr., O. (2004). Furtado-Gallagher movement skill assessment: collecting evidence for content-related validity. Retrieved July 14, 2008, from http://etd.library.pitt.edu/ETD/available/etd-12042004-093510/.

Gabbard, C. (2007). *Lifelong motor development* (5th ed.). San Francisco, CA: Benjamin Cummings.

Gallahue, D., & Donnelly, C. (2003). *Developmental physical education for all children*. Champaign, IL: Human Kinetics.

Gallahue, D., & Donnelly, C. (2007). *Developmental Physical Education for All Children*. Champaign, IL: Human Kinetics.

Gallahue, D. L., & Ozmun, J. C. (2002). *Understanding motor development : infants, children, adolescents, adults*. Dubuque, IA: McGraw-Hill.

Graham, G., Holt/Hale, S. A., & Parker, M. (2007). *Children moving : a reflective approach to teaching physical education*. New York: McGraw-Hill.

Halverson, L., & Williams, K. (1985). Developmental Sequences for Hopping Over Distance: A Prelongitudinal Screening. *Research Quarterly for Exercise and Sport*, *56*(1), 37-44.

Harper, C. J., & Struna, N. L. (1973). *Case Studies in the Development of One-Handed Striking.* Minneapolis, MN: American Association for Health, Physical Education and Recreation.

Haywood, K., & Getchell, N. (2009). *Life Span Motor Development*. Champaign, IL: Human Kinetics.

Haywood, K., & Getchell, N. (2005). *Life span motor development*. Champaign, Ill.; Leeds: Human Kinetics.

Hume, C., Okely, A., Bagley, S., Telford, A., Booth, M., Crawford, D., et al. (2008). Does weight status influence associations between children's fundamental movement skills and physical activity? *Research Quarterly for Exercise and Sport*, *79*(2), 158-165.

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kelso, S., & Clark, J. (1982). *The Development of movement control and coordination*. Wiley series in developmental psychology. New York: Wiley.

Loovis, E. M., & Ersing, W. F. (1979). *Assessing and Programming Gross Motor Development for Children*. Loudonville, OH: Mohican Textbook Pub.

Mazzardo, O. (2008). The Relationship of Fundamental Movement Skills and Level of Physical Activity in Second Grade Children. *Electronic Dissertation of University of Pittsburgh*. Retrieved June 30, 2008, from http://etd.library.pitt.edu/ETD/available/etd-04302008-144205/.

McClenaghan, B. A., & Gallahue, D. (1978). *Fundamental movement: a developmental and remedial approach*. Philadelphia, PA: Saunders.

Mills, C. (1983). A Comparison of Three Methods of Establishing Cut-Off Scores on Criterion-Referenced Tests. *Journal of Educational Measurement*, *20*(3), 283-92.

NASPE. (2004). *Moving into the future: national standards for physical education*. Reston, VA: McGraw-Hill.

Okely, A., Booth, M., & Chey, T. (2004). Relationships between Body Composition and Fundamental Movement Skills among Children and Adolescents. *Research Quarterly for Exercise and Sport*, *75*(3), 238.

Okely, A., Booth, M., & Patterson, J. W. (2001). Relationship of physical activity to fundamental movement skills among adolescents. *Medicine & Science in Sports & Exercise*, *33*(11), 1899.

Payne, V. G., & Isaacs, L. D. (2008). *Human motor development : a lifespan approach*. Boston, MA: McGraw-Hill.

Rarick, L. (1982). Descriptive research and process-oriented explanations of the motor development of children. In S. Kelso & J. Clark (Eds.), *The Development of movement control and coordination*, Wiley series in developmental psychology. Chichester [West Sussex]; New York: Wiley.

Roberton, M. A. (1982). Describing 'stages' within and across motor tasks. In S. Kelso & J. Clark (Eds.), *The Development of movement control and coordination*, Wiley series in developmental psychology. Chichester [West Sussex]; New York: Wiley.

Roberton, M. A. (1989). Developmental sequence and developmental task analysis. In J. S. Skinner (Ed.), *Future directions in exercise and sport science research*. Champaign, IL: Human Kinetics.

Roberton, M. A. (1977). Stability of stage categorizations across trials: Implications for the stage theory of overarm throw development. *Journal of Human Movement Studies*, *3*, 49-59.

Roberton, M. A., & Halverson, L. E. (1984). *Developing children and their changing movement: a guide for teachers*. Philadelphia, PA: Lea & Febiger.

Roberton, M. A., Williams, K., & Langendorfer, S. (1980). Pre-longitudinal screening of motor development sequences. *Research Quarterly for Exercise and Sport*, *51*(4), 724-31.

Runion, B. P., Roberton, M. A., & Langendorfer, S. (2003). Forceful Overarm Throwing: A Comparison of Two Cohorts Measured 20 Years Apart. *Research Quarterly for Exercise & Sport*, *74*, 324-330.

Safrit, M. J., & Wood, T. M. (1995). *Introduction to Measurement in Physical Education and Exercise Science*. St. Louis, MI: Mosby.

Sayre, N., & Gallagher, J. (2001). *The young child and the environment: issues related to health, nutrition, safety, and physical activity*. Boston, MA: Allyn and Bacon.

Seefeldt, V., & Haubenstricker, J. (1976). Developmental Sequences of Fundamental Motor Skills. *Unpublished research, Michigan State University*.

Seefeldt, V., & Haubenstricker, J. (1982). Patterns, phases, or stages: an analytical model for the study of developmental movement. In S. Kelso & J. Clark (Eds.), *The Development of movement control and coordination*, Wiley series in developmental psychology. Chichester: New York: Wiley.

Spitzer, R., & Fleiss, J. (1974). A re-analysis of the reliability of psychiatric diagnosis. *The British journal of psychiatry: the journal of mental science*, *125*(0), 341-7.

Taylor, J. L. (1979). Development of the physical education observational instrument using generalizability study theory. *Research Quarterly*, *50*(3), 468-481.

Thelen, E. (2000). Motor development as foundation and future of developmental psychology. *International Journal of Behavioral Development*, *24*(4), 385-397.

Thorn, D. W., & Deitz, J. C. (1989). Examining content validity through the use of content experts. *The Occupational Therapy Journal of Research*, *9*(6), 334-347.

Ulrich, D., & Branta, C. (1988). Developmental Gross Motor Skill Ratings: A Generalizability Analysis. *Research Quarterly for Exercise and Sport*, *59*(3), 203-09.

Ulrich, D., & Sanford, C. B. (2000). *Test of Gross Motor Development: examiner's manual*. TX: Pro-Ed.

Wickstrom, R. L. (1983). *Fundamental Motor Patterns* (Third.). Philadelphia, PA: Lea & Febiger.

Zhu, W., & Cole, E. L. (1996). Many-Faceted Rasch Calibration of a Gross Motor Instrument. *Research Quarterly for Exercise and Sport*, *67*(1), 24-34.