Psychometric properties of two systematic observation tools, the System for Observing Fitness Instruction Time (SOFIT) and the Children’s Activity Rating Scale (CARS), were examined for use with individuals with mental retardation (MR). Eleven children with MR were videotaped while participating in gym-based physical activity. Accelerometer data were collected and synchronized with the observational data. Three raters coded each videotape twice each with SOFIT and CARS.

Generalizability theory analysis indicated that SOFIT had low error variance due to rater, trial, and interaction terms. This provides evidence that SOFIT has good reliability (Φ=0.98). Concurrent validity evidence for SOFIT indicated that SOFIT may not be an appropriate tool for use with individuals with MR. Validity coefficients (r) between accelerometer data and SOFIT scores ranged from -0.44 to +0.39, indicating less than 20% shared variability.
G-theory analysis for CARS also indicated that CARS demonstrates sufficient reliability for use with individuals with MR (Φ=0.76). There was a higher level of error variance associated with rater for the CARS instrument which indicates that more training on this tool may be necessary. Validity evidence for CARS was somewhat stronger than SOFIT with correlations between accelerometer data and CARS interval scores ranging from -0.52 to +0.79 ($r^2=0.62$).

Systematic observation tools may not differentiate between the low physical activity levels of individuals with mental retardation with may have caused the low validity levels. These findings indicate that systematic observations tools provide strong reliability evidence and weak validity evidence for use with individuals with mental retardation, and should not be used for this group.
Psychometric Properties of Two Systematic Observation Techniques for Assessing Physical Activity Levels in Children with Mental Retardation

by
Christina Anne Taylor

A THESIS
Submitted to
Oregon State University

in partial fulfillment of the requirements for the degree of
Master of Science

Presented May 2, 2003
Commencement June 2003

APPROVED:

Redacted for privacy

Major Professor, representing Movement Studies in Disability

Redacted for privacy

Chair of Department of Exercise and Sport Science

Redacted for privacy

Dean of the Graduate School

I understand that my thesis will become a part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Redacted for privacy

Christina Anne Taylor, Author
I would like to thank the members of my committee. To Dr. Robert Nye for saying “YES” when others had said “NO” and filling the much needed role of graduate representative. To Dr. Jeff McCubbin for accepting me into this program and providing me with so many opportunities for “hands-on” learning. To Dr. Terry Wood, for opening my eyes to the world of measurement! And to Dr. Joonkoo Yun, for being a constant source of motivation since I arrived at Oregon State University and for always asking how I am before asking how my work is. Thank you for challenging me. I am extremely grateful for your patience with me throughout my thesis work.

To Jason Russell and Soyeun Kim, for the hours you spent filming and coding tapes. This project would not have happened without your commitment and hard work. I appreciate every one of the 3300 intervals you coded.

To Keith Johnston, thank you for your help with data collection the first hectic week, and for your support through every step of my thesis. It’s nice to know there are other people in the same place in life!

To “my girls”, Jo Swinemeyer, Alison Bigelow, and Erika Therrien, the best friends a person could ask for, you have supported me even when I’m 4000 miles from home. To those who have touched my life here at Oregon State University: Kelly Billon, Scott Fortner, Kris Hagar, Jim Tabb, and Dan Tindall. Thank you for helping me call Corvallis home, and teaching me what it means to be a Beaver.
To my family: Tanya, Mark, Charles, Andy, Sheila, Chris, and Dave, for letting me go through all my “phases” with you. I’m proud to be your little (or big) sister and cherish the rare times we are all together. To my late mom, for teaching me it’s okay to cry sometimes, and to sing… no matter what.

And finally, to my Dad, for teaching me to balance work and play, for your constant support and encouragement, and most of all for your daily e-mail after the evening news. You’ll never know how much those e-mails have meant to me as I have gone through this process. I love you so much.
CONTRIBUTION OF AUTHORS

Dr. Joonkoo Yun was involved in the data analysis, and writing of the manuscript.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>1</td>
</tr>
<tr>
<td>RATIONALE</td>
<td>4</td>
</tr>
<tr>
<td>CHAPTER 2: PSYCHOMETRIC PROPERTIES OF TWO SYSTEMATIC OBSERVATION TECHNIQUES FOR ASSESSING PHYSICAL ACTIVITY LEVELS IN CHILDREN WITH MENTAL RETARDATION</td>
<td>9</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>10</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>11</td>
</tr>
<tr>
<td>METHODS</td>
<td>14</td>
</tr>
<tr>
<td>RESULTS</td>
<td>24</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>29</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>36</td>
</tr>
<tr>
<td>CHAPTER 3: SUMMARY</td>
<td>39</td>
</tr>
<tr>
<td>RESEARCH CONCLUSIONS</td>
<td>39</td>
</tr>
<tr>
<td>FUTURE RESEARCH DIRECTIONS</td>
<td>41</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>43</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>49</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>Descriptive Statistics for Children's Percent MVPA for SOFIT</td>
</tr>
<tr>
<td>2</td>
<td>Descriptive Statistics for Children's Activity Scores for CARS</td>
</tr>
<tr>
<td>3</td>
<td>Variance Component Estimates and their Relative Magnitudes for SOFIT</td>
</tr>
<tr>
<td>4</td>
<td>Variance Component Estimates and their Relative Magnitude for CARS</td>
</tr>
<tr>
<td>5</td>
<td>Within Subject Correlations (r) for SOFIT and CARS</td>
</tr>
<tr>
<td>6</td>
<td>Decision Study for SOFIT (p x r x t Design)</td>
</tr>
<tr>
<td>7</td>
<td>Decision Study for CARS (p x r x t Design)</td>
</tr>
<tr>
<td>Appendix</td>
<td>Title</td>
</tr>
<tr>
<td>----------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>A</td>
<td>REVIEW OF LITERATURE</td>
</tr>
<tr>
<td>B</td>
<td>INFORMED CONSENT DOCUMENT</td>
</tr>
<tr>
<td>C</td>
<td>INSTITUTIONAL REVIEW BOARD APPROVAL</td>
</tr>
<tr>
<td>D</td>
<td>SAS COMMAND FILE FOR G-THEORY ANALYSIS</td>
</tr>
<tr>
<td>E</td>
<td>SOFIT CODING FORM</td>
</tr>
<tr>
<td>F</td>
<td>CARS CODING FORM</td>
</tr>
</tbody>
</table>
DEDICATION

This thesis is dedicated to my father

L.T. Taylor

for giving me both roots and wings.
CHAPTER 1: INTRODUCTION

BACKGROUND

The relationship between physical activity and the reduction of health risks has been well-documented in research literature (Blair et al., 1995; Brosse, Sheets, Lett, & Blumenthal, 2002; Hassmen, Koivula & Uutela, 2000; Paffenbarger, & Lee, 2000). Healthy People 2010 named physical activity as one of the 10 leading health indicators which “illuminate individual behaviors, physical and social environmental factors, and important health system issues that greatly affect the health of individuals and communities” (United States Department of Health and Human Services [USDHHS], 2000). “Physical Activity and Health: A Report of the Surgeon General” (Centers for Disease Control and Prevention [CDC], 1996a) suggests that people of all ages can benefit from physical activity. Health benefits of regular physical activity include an increase of immune functioning and longevity, and a reduced risk of cardiovascular disease, adult onset diabetes, obesity, hypertension, osteoporosis, and some types of cancer (Blair et al., 1995; Blair et al., 1989; Helmrich, Ragland, Leung, Paffenbarger, 1991; Helmrich,
Psychological benefits of regular participation in physical activity have not been studied extensively but preliminary studies have addressed the impact of physical activity on depression, well-being, anxiety, and self-efficacy (Brosse, Sheets, Lett, & Blumenthal, 2002). Brosse et al. (2002) indicated that there is considerable support that regular physical activity has a positive impact on depression. Hassmen, Koivula, and Uutela (2000) indicated regular exercisers experienced significantly less depression, anger, cynical distrust, and stress than sedentary individuals.

The association between physical activity and physical fitness has been demonstrated in the literature (Pate, Dowda, & Ross, 1990). Individuals who participate in regular physical activity are more likely to have higher physical fitness ratings. This is of particular importance when trying to increase physical activity levels in children with mental retardation (MR) because disparities in physical fitness exist between individuals with and without MR. When compared with their peers without MR, individuals with MR have poor physical fitness levels. The average deficit in cardiovascular fitness across several studies is 25 to 30% when compared with expected values for individuals without disabilities. Individuals with MR have a greater prevalence of obesity and lower muscular strength and endurance than individuals without MR (Fernhall, 1993).
Eliminating health disparities between individuals with and without disabilities is one of the goals of Healthy People 2010 (United States Department of Health and Human Services [USDHHS], 2000). Increasing physical activity levels in individuals with disabilities is one way to help eliminate disparities. Instruction to ensure proper motor skill development and participation in lifetime physical activities early in life will contribute to opportunities for active lifestyles later in life (Corbin, Pangrazi, & Welk, 1994).

Information about physical activity levels is important in health-related research, supervision, and curriculum development (McKenzie, 1991). Recent research has used many techniques to attempt to quantify physical activity levels in both children and adults. Self-report methods, heart rate monitors, pedometers, accelerometers, the doubly labeled water technique, and systematic observation have all been used (Rowe, Schuldheisz, & van der Mars, 1997).

Systematic observation is a process used to record behavior based on specifically designed guidelines and coding procedures (van der Mars, 1989). Systematic observation can be a reliable and objective observational technique. Tools can be used to record information pertaining to observable behaviors, events, and episodes. Observations may be made based on form of skills or outcomes and consequences of behavior (van der Mars, 1989). Event recording, duration recording, interval recording, and momentary time sampling are tactics used to code (record) behaviors with systematic observation tools. The tactic selected will
depend on what information is to be obtained from the results, and the duration and repeatability of the behaviors to be coded (van der Mars, 1989).

There are many advantages to using systematic observation including (a) observation can take place in a natural activity setting rather than a research lab (Epstein, McGowan, & Woodall, 1984); (b) systematic observation requires very little equipment and is inexpensive (Puhl, Greaves, Hoyt, & Baranowski, 1990); and (c) it is a direct measure of behavior, systematic observation is sometimes used as a criterion measure to validate other methods of measuring physical activity. Limitations of systematic observation include potential for participant reactivity (participants altering their behavior because of the observer/camera presence), and its limitation to observable behaviors and descriptive information (van der Mars, 1989).

Many systematic observation tools have been developed to measure levels of physical activity (McKenzie, Sallis, & Nader, 1991; O’Hara, Baranowski, Simons-Morton, Wilson, & Parcel, 1989; Puhl et al., 1990). Information obtained through systematic observation may include the type, intensity, duration, physical environment, and social environment of activity (O’Hara et al., 1989).

RATIONALE

Systematic observation tools for determining physical activity levels are useful for both researchers and practitioners. McKenzie (1991) stated “insufficient
information exists to indicate which systems are most accurate and which can be used most reliably... Simultaneous coding of children's physical activity from videotapes using different observation systems is needed” (p.27). Furthermore, there is very limited information regarding the appropriateness of using systematic observation for individuals with mental retardation.

The Children's Activity Rating Scale (CARS) was developed by Puhl et al. (1990) to categorize activities of different intensities, and estimate energy expenditure levels. Although it appears to be an appropriate tool to assess physical activity levels for children with MR the psychometric properties of CARS have not been examined for individuals with MR.

The System for Observing Fitness Instruction Time (SOFIT) was developed by McKenzie et al. (1991) to provide instructors with information about students' activity levels, teacher behaviors, and curriculum content variables. SOFIT has been used in studies with adults with mental retardation with an assumption made that it is appropriate for this population (Stanish, 1999).

Hodge (2001) conducted a validation study for SOFIT using a small sample of children with mental retardation and results indicated it may be an appropriate tool for use with children with MR. Heart rate monitors were used as the criterion measure in this study and were correlated with SOFIT scores. There were; however, some methodological issues with Hodge's design. First, the heart rate monitor intervals and SOFIT intervals were not synchronized. Second, no reliability information was cited. Third, Hodge did not use a conventional
correlation method. Heart rate values and SOFIT levels for each person were correlated, therefore; the correlations cited were for each individual rather than the sample of study. This is a descriptive measure and generalization of her findings can be limited.

Hodge (2001) suggested that SOFIT has face validity for individuals with mental retardation because there are no physical disabilities that would affect posturing (sitting, lying down) and subsequently scoring of this instrument. Individuals with mental retardation; however, have more sedentary lifestyles than individuals without mental retardation (Deener & Horvat, 1995; Fernhall, 1993; Fernhall et al., 1996) and it is not known if systematic observation will be able to discriminate among activity levels for this population. Yun and Ulrich (2002) argued that validity generation should be limited to populations with similar behavioral characteristics. Therefore, there is a need to evaluate the psychometric properties of systematic observation tools with individuals with mental retardation to examine if reliability and validity evidence can be generalized to this population. The ability to use systematic observation tools with individuals with mental retardation would provide researchers and practitioners a way to measure activity levels, and to compare activity levels between individuals with and without mental retardation.

The purpose of this study was to examine and compare the psychometric properties of two physical activity systematic observation systems for use with children with MR. Specifically, generalizability of scores over raters and trials was
examined along with criterion related evidence for validity using accelerometer data.

Research Questions

The following research questions were investigated in this study:

1) How much of the error variance in activity level scores is due to the rater?
2) How much of the variance in activity level scores is due to the trial?
3) Do systematic observation techniques have high intrarater reliability?
4) Do systematic observation techniques have high interrater reliability?
5) Do SOFIT and CARS demonstrate concurrent validity when correlated with accelerometer data?
6) What is the optimal measurement protocol for SOFIT and CARS?

Assumptions

1) The Actiwatch accelerometer captured children's physical activity accurately.
2) SOFIT intervals of 15 seconds are appropriate.
Delimitations

1) Data were collected during the Special Motor Fitness Clinic in Winter Term, 2003.

2) Only the psychometric properties of SOFIT and CARS were examined.

3) Only pre-selected students were participants in the study (e.g. Subjects were not randomly sampled).
CHAPTER 2

Psychometric Properties of Two Systematic Observation Techniques for Assessing Physical Activity Levels in Children with Mental Retardation

Christina Anne Taylor
ABSTRACT

Purpose: To examine the psychometric properties of two physical activity systematic observation instruments for use with individuals with mental retardation (MR).

Methods: Eleven children with MR were videotaped while participating in gym-based physical activity. Three raters coded the data twice each with the System for Observing Fitness Instruction Time (SOFIT), and the Children's Activity Rating Scale (CARS). Accelerometer data was synchronized with the coding intervals and used as the criterion measure for concurrent validity.

Results: SOFIT and CARS both demonstrated high generalizability (Φ=0.98 and 0.75) across raters and trials. SOFIT demonstrated low concurrent validity (r = -0.44 to +0.39). CARS demonstrated low/moderate concurrent validity (r = -0.52 to +0.79).

Conclusions: SOFIT may be administered with one rater and one trial with high reliability evidence (Φ=0.94). CARS may be administered with one trial and four raters to achieve generalizability (Φ=0.80). CARS raters require more training than provided in this study. Based on the validity evidence in this study, SOFIT is inappropriate for use with individuals with MR. CARS should be re-evaluated with additional rater training to determine if validity evidence improves.
INTRODUCTION

Regular participation in physical activity has many benefits including an increase in immune functioning and longevity, and a reduction in cardiovascular disease, adult onset diabetes, obesity, osteoporosis, some types of cancer, depression, and stress (Blair et al., 1995; Blair et al., 1989; Hassmen, Koivula, & Uutela, 2000; Helmrich, Ragland, Leung, & Paffenbarger, 1991; Helmrich, Ragland, & Paffenbarger, 1994; Paffenbarger, 2000; Paffenbarger, Wing, & Hyde, 1978; Sesso, Paffenbarger, Ha, & Lee, 1999; Sesso, Paffenbarger, & Lee, 2000). Healthy People 2010 set two goals aimed to improve the quality of life for individuals with disabilities. The first goal was to eliminate health disparities between individuals with and without disabilities. The second was to reduce the frequency of secondary conditions in individuals with disabilities (United States Department of Health and Human Services [USDHHS], 2000).

Health disparities and secondary conditions are both concerns for individuals with mental retardation (MR). Individuals with MR are at a greater risk for poor cardiovascular health and obesity due to poor physical activity habits, and sedentary lifestyle (Deener & Horvat, 1995; Fernhall, 1993; Fernhall et al, 1996). Because of the many benefits of physical activity, increasing physical activity levels in individuals with MR may help eliminate some of these health disparities. In order to provide appropriate physical activity services to individuals with MR, accurately evaluating one's level of physical activity in a non-research (clinical)
setting becomes an important first step in planning and developing physical intervention programs for children with MR. Methods of measuring physical activity levels in field-based assessment should not only demonstrate good psychometric properties, but also have low cost and be easy to use in practical situations. Accurate information about physical activity levels obtained from a field-based method is important for health-related research, supervision, and curriculum development (McKenzie, 1991).

Systematic observation is a field-based assessment method that is used to record behavior based on specifically designed guidelines and coding procedures (van der Mars, 1989). Information obtained through systematic observation may include the type, intensity, duration, physical environment, and social environment of activity (O’Hara, Baranowski, Simons-Morton, Wilson, & Parcel, 1989). Many systematic observation tools including the System for Observing Fitness Instruction Time (SOFIT), Children’s Physical Activity Form, Movement of Arms and Legs, Activity Patterns and Energy Expenditure, and the Children’s Activity Rating Scale (CARS), have been developed to measure levels of physical activity in clinical settings for children without disabilities (Epstein, McGowan, & Woodall, 1984; Hovell, Bursick, Sharkey, & McClure, 1978; McKenzie, Sallis & Nader, 1991; O’Hara et al., 1989; Puhl, Greaves, Hoyt, & Baranowski, 1990). However, McKenzie (1991) stated “insufficient information exists to indicate which systems are most accurate and which can be used most reliably... Simultaneous coding of children’s physical activity from videotapes using different observation systems is
needed" (p.27). Furthermore, there is limited information regarding the appropriateness of using systematic observation for individuals with MR.

Different coding techniques have been proposed for use with systematic observation tools. SOFIT is an interval and momentary time sampling instrument, while CARS is a partial interval coding instrument. Both instruments provide methods for calculating a numerical summary score which could be used in data analysis. These instruments were selected for psychometric analysis in this study because of their different approaches to quantifying physical activity.

Although both tools have demonstrated acceptable psychometric properties for children without disabilities, no psychometric properties have been established for CARS with individuals with MR, and limited information regarding validity of SOFIT for individuals with MR has been reported (Hodge, 2001). Hodge cited concurrent validity correlations ($r = 0.06$ and $0.90$) between heart rates and SOFIT scores for eight children with MR. However, there are methodological issues with Hodge’s design. First, the heart rate monitor intervals and SOFIT intervals were not synchronized. Second, no reliability information was cited. Third, Hodge did not use a conventional correlation method. Heart rate values and SOFIT levels for each person were correlated rather than group data; therefore, the correlations reported are for each individual. This is a descriptive measure and may not be generalized.

Hodge (2001) suggested that SOFIT has face validity for individuals with MR because there are no physical disabilities that would affect posturing (sitting,
lying down) and subsequently scoring of this instrument. However, considering individuals with MR have more sedentary lifestyles than individuals without MR (Deener & Horvat, 1995; Fernhall, 1993; Fernhall et al., 1996), it is not known if systematic observation will be able to discriminate among activity levels for this population. Yun and Ulrich (2002) suggested that validity generation should be limited to populations with similar behavioral characteristics. Therefore, there is a need to evaluate the psychometric properties of systematic observation tools with individuals with MR to examine if reliability and validity evidence can be generalized to this population. The ability to use systematic observation tools with individuals with MR would provide researchers and practitioners a way to measure activity levels, and to compare activity levels between individuals with and without MR. The purpose of this study is to compare two systematic observation systems to examine their psychometric properties for use with individuals with mental retardation.

METHODS

Participants

Participants in this study were 11 children between the ages of 6 and 14 years \(M = 10.5, SD = 2.50\) years previously diagnosed with MR, who participated in a university-based community out-reach program. All participants of this
The Children's Activity Rating Scale (CARS), System for Observing Fitness Instruction Time (SOFIT), and Actiwatch accelerometers were used to measure physical activity levels of participants.

CARS was developed to categorize physical activities of different movement intensities (Puhl et al., 1990). These levels range from no movement to very fast translocation. An activity rating score was calculated for each minute by taking the mean of the activity levels recorded in the interval. Puhl et al. (1990) reported interrater agreement of 84.1% (SD=10.1%) for CARS. This study examined the mean percent agreement over 389 paired observations (13,651
intervals. Validity evidence for CARS levels was investigated using VO₂ uptake and coding levels. Significant differences (p<0.05) in the VO₂ uptake values for the five CARS levels demonstrate activity levels may be differentiated using the CARS system.

SOFIT is a momentary time sampling and interval recording systematic observation tool that divides physical activity behaviors into five categories ranging from lying down to very active. Levels 4 and 5 are equivalent to moderate to vigorous physical activity (MVPA) (Rowe, Schuldheisz, & van der Mars, 1997). An activity score can be computed using the percent of the intervals in which the individual was performing MVPA. Behaviors are coded according to what the participant is doing at the moment the interval ends.

SOFIT has proven psychometric properties for individuals without disabilities. Rowe et al. (1997) reported stability reliability across tasks ranging from R=0.80 for jogging to R=0.91 for curl-ups. These intraclass correlations were calculated using the mean heart rate over two days. The time lag between the two measurement occasions varied from one day to one week. Intraclass correlations for heart rates of individuals measured on only one day were calculated and ranged from R=0.66 for jogging to R=0.84 for curl-ups.

Validity evidence in the Rowe et al. (1997) study was examined using individuals' heart rates across the different SOFIT levels. A repeated measures analysis of variance (ANOVA) was employed. There was a significant task-effect (p<0.001) indicating that there are significant differences in the mean heart rates by
task (SOFIT level). Post-hoc analysis using Scheffe’s contrast showed significant differences in the mean heart rates for all coding levels of SOFIT except lying down and sitting. SOFIT demonstrated high interrater reliability determined using intraclass correlation after training of 0.88 for the student activity portion (McKenzie et al., 1991).

Actiwatch (Minimitter, 2000) accelerometers were used as the criterion measure for physical activity levels for comparison to CARS and SOFIT data. Accelerometers provide data in the form of activity counts. The activity count represents the summed magnitude of the accelerations. This gives an indication of how much the individual moved in the interval recorded. Nichols, Morgan, Sarkin, Sallis, and Calfas (1999) reported 90% of the variance in energy expenditure was explained by the accelerometer activity counts.

Procedures

Data collection involved videotaping physical activity of participants during planned program activities. Upon arrival at the gym an accelerometer was placed on the non-dominant hand of each participant. Hand placement was selected because hip placement tends to underestimate energy costs of physical activities (Strath, Bassett, Swartz, & Thompson, 2001). Accelerometers were set to collect a movement count every 15 seconds. Hand dominance was determined one of four ways: a) asking the child which hand they write with, b) giving the child a pen and
asking them to write their name on a piece of paper, c) asking their parent, or d) asking their clinician. Each participant was videotaped for 15 minutes while participating in gym-based physical activity. Videotaping took place during a variety of lesson contexts. The program follows a four-part lesson plan of an introductory activity, a fitness activity, a lesson focus, and a game. Filmed activities included in filming were hockey, lummi sticks, parachute, fitness activities, rhythmic gymnastics, and disc throwing.

Coding of the videotapes using SOFIT and CARS was performed by three students. Raters were provided with approximately 7 hours of training for the two systematic observation tools (CARS and SOFIT). Raters also practiced coding independently with multiple practice tapes prior to actual data collection. Training on each system followed the protocol outlined by van der Mars (1989).

Prior to coding, interrater reliability among raters was established. For CARS, interrater reliability was calculated using the percent agreement for each interval among raters. For example, if for the first interval, rater A coded levels 2 and 3, and rater B coded only level 2, then the percent agreement is 1/2, or 0.5 (Puhl et al., 1990). Mean percent agreement for all intervals was then used as the estimate of interrater reliability. Mean percent agreement between raters after training ranged from 87% to 93%.

Interrater reliability for SOFIT was estimated using a strict interval by interval agreement procedure (McKenzie et al., 1991). The number of intervals agreed upon was divided by the total number of intervals to obtain an estimate of
interrater reliability. After training, the interrater reliability among the three raters was 0.90. Also, each rater’s performance was compared with an expert who had used SOFIT for 9 years and had conducted validity studies for the instrument. The interrater reliability between the expert and the raters prior to data analysis ranged from 0.83 to 0.87.

Once coding began, each tape was coded four times, twice each with SOFIT and CARS. SOFIT was coded using 15 second intervals to be synchronized with accelerometer data. CARS was coded using one minute intervals and was also synchronized with one minute summed accelerometer counts. A minimum 1 week delay was observed between trials. The coding of tapes was counterbalanced to reduce observer drift. Observer drift in systematic observation is when individuals change coding rules and interpret categories differently than the proper rules (van der Mars, 1989). Observer drift was a concern in this study because raters were switching between instruments frequently. The counterbalancing was set up as follows: (a) Rater A always coded SOFIT and then CARS; (b) Rater B always coded CARS and then SOFIT; and (c) Rater C coded SOFIT and then CARS the first time a subject was coded and coded with CARS and then SOFIT the second time a subject was coded. Counterbalancing ensures that SOFIT was coded before CARS as many times as CARS was coded before SOFIT.
Pilot Studies

A pilot study was performed to familiarize researchers with the equipment, determine optimal camera placement, and reduce subject reactivity when the actual data collection began. Compatible camera placement was a location where the whole activity space was visible, cameras were not limiting the programming activity space, and cameras were relatively discrete. The cameras remained in the same location every week and a camera was present in the gym on at least one occasion prior to any participant being filmed to reduce participant reactivity.

Data Analysis

Generalizability analysis. Generalizability theory (G-theory) is a data analysis technique which allows researchers to examine sources of variability in data. Traditional reliability analysis provides only a single score to evaluate the reliability of measurement. G-theory not only provides a single coefficient, but also it provides magnitudes of error from multiple sources (Shavelson & Webb, 1991). G-theory focuses on the magnitude of variability related to the participants, possible sources of error (facets), and their interactions. When researchers are able to pinpoint the sources of error, it provides them with information needed to improve the reliability of the measurement.
In this study, a two-facet generalizability study design was used to examine the sources of variability in assessing physical activity levels. For SOFIT, the percent of time spent in MVPA was used for the physical activity level. MVPA is defined as physical activity levels in SOFIT of 4 (walking) or 5 (running) (Rowe et al., 1997). For CARS, the mean interval score was used for the activity score. No conversion to MVPA was available for CARS. The data were analyzed separately for each instrument using a completely crossed 2 by 3 (trail by rater) ANOVA with all participants' activity coded on two trials, by three raters. Seven sources of variability were estimated from the ANOVA results. The variance associated with participant ($\sigma^2_p$), trial ($\sigma^2_t$), rater ($\sigma^2_r$), three two-way interactions, and the residual term (three-way interaction plus error) were determined using the VARCOMP procedure from SAS, version 8 (SAS Institute, n.d.). Negative variance components were set to zero in subsequent calculations as suggested by (Morrow, 1989).

The participant variance component ($\sigma^2_p$), indicates how much the children differed in their level of physical activity during program activities. It is assumed that the majority of the variability in scores should be due to the true differences among the participants being observed. Sampling variability due to trial ($\sigma^2_t$) gives an indication of sources of error due to the test-retest reliability of the raters. Sampling variability due to the rater ($\sigma^2_r$), gives an indication of inter-rater reliability. Three two-way interaction terms were calculated. The person-by-trial ($\sigma^2_p \times \sigma^2_t$) interaction gives an indication of how the trials were influenced by the
individual activity patterns (i.e. were between participant differences consistent across trials), since both trials were captured on videotape of the child this interaction term should be close to zero in this study. Variability due to person-by-rater interaction ($\sigma^2_p \times \sigma^2_r$) provides information about if rater's coding is influenced by individual activity patterns. The rater-by-trial ($\sigma^2_{ri} \times \sigma^2_t$) interaction provides information about how the raters are influenced by coding the videotapes on two occasions (i.e. were rater differences consistent across trials). The residual variance component (three-way interaction term) will include systematic and random error (Shavelson & Webb, 1991). Percent variances were calculated by dividing the variance for each component by the total variance.

Once variance components were computed, generalizability coefficients (g-coefficients) were calculated to give an indication of reliability. G-coefficients give an indication of how sound the generalization for an individual's observed score is to their true universe score. G-coefficients are calculated differently depending on what sort of decisions are to be made from the results. If interpretations pertain to the relative standing of participants then a relative decision is made. G-coefficients for relative decisions are calculated using all the interactions terms between the facets and the object of measurement. If decisions are based on the absolute score for participants and relative standing has no bearing then absolute decisions are made. In the case of absolute decisions all variance components except the object of measurement are used in the calculation of an index of dependability. The index of dependability is the absolute decision equivalent of the G-coefficient in relative
decisions (Shavelson & Webb, 1991). Because relative standing of the individuals in the study had no bearing on the results, estimates for measurement error were based on absolute interpretations.

**Validity.** Validity coefficients were calculated using Pearson product moment correlations between results from Actiwatch accelerometers and the results from the two systemic observation tools. The nature of these data can be problematic for correlations. When you have many intervals from the same subject it may lead to loss of independent observations. However, if you take the mean values for each subject in may lead to a loss of variability in the data. Although unconventional, validity coefficients were calculated by three methods, each method was selected because of its specific measurement properties.

The first method correlated interval scores on the observation tool for each subject with the corresponding interval accelerometer counts, providing a range of within-subject correlations. The observation tool interval score used was the mean rating across six trials (3 raters X 2 trials) for each interval. For SOFIT, 60 intervals went into each subject's correlation because of the 15 seconds interval length. For CARS, 15 intervals went into each subject’s correlation because of the 1-minute interval length. This method does not violate the assumption of independence of measurement but it also may not provide generalizable results.

The second method used an overall activity score for each subject and correlated it with the total accelerometer count for each subject. For SOFIT, the mean percent MVPA from the six trials (3 raters X 2 trials) was used for the overall
activity score for each subject. For CARS, the mean of all interval activity scores across the six trials was used for the subject’s overall activity score. Using mean scores in this method loses information about individual variability and may underestimate the validity coefficient because it will reduce the variability of scores. The advantage of this method is that it may provide more generalizable results.

The third method correlated all interval observation scores for all subject intervals and all accelerometer counts. This method would also violate the assumption of independence of measurement but be more generalizable than the first method. The advantage of this method is that it maximizes the variability of scores.

RESULTS

According to SOFIT, children with MR participated in MVPA an average of 26.48% (SD=14.77%) of the programming time. Table 1 summarizes each child’s average percent MVPA across the six trials during the physical activity program.
### Table 1
Descriptive Statistics for Children’s Percent MVPA for SOFIT

<table>
<thead>
<tr>
<th>Subject</th>
<th>M</th>
<th>SD</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52.97</td>
<td>4.99</td>
<td>0.094</td>
</tr>
<tr>
<td>2</td>
<td>29.93</td>
<td>2.94</td>
<td>0.098</td>
</tr>
<tr>
<td>3</td>
<td>34.85</td>
<td>5.61</td>
<td>0.161</td>
</tr>
<tr>
<td>4</td>
<td>36.39</td>
<td>3.23</td>
<td>0.089</td>
</tr>
<tr>
<td>5</td>
<td>14.38</td>
<td>1.43</td>
<td>0.099</td>
</tr>
<tr>
<td>6</td>
<td>43.6</td>
<td>2.86</td>
<td>0.066</td>
</tr>
<tr>
<td>7</td>
<td>30.58</td>
<td>1.50</td>
<td>0.049</td>
</tr>
<tr>
<td>8</td>
<td>5.6</td>
<td>1.16</td>
<td>0.207</td>
</tr>
<tr>
<td>9</td>
<td>10.83</td>
<td>2.04</td>
<td>0.188</td>
</tr>
<tr>
<td>10</td>
<td>15.18</td>
<td>1.51</td>
<td>0.099</td>
</tr>
<tr>
<td>11</td>
<td>16.92</td>
<td>7.53</td>
<td>0.445</td>
</tr>
<tr>
<td>All subjects</td>
<td>26.48</td>
<td>14.77</td>
<td>0.558</td>
</tr>
</tbody>
</table>

*Note: Means are based on the average score across the three raters (six trials).*

CARS indicates that the overall mean activity level of the children was 2.37 (SD=0.35) on a scale of 1 to 5 where level 1 is stationary with no limb movement and level 5 is very fast translocation. Table 2 summarizes each child’s physical activity across three raters and two trials as coded by CARS.

### Table 2
Descriptive Statistics for Children’s Activity Scores for CARS

<table>
<thead>
<tr>
<th>Subject</th>
<th>M</th>
<th>SD</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.84</td>
<td>0.24</td>
<td>0.084</td>
</tr>
<tr>
<td>2</td>
<td>2.25</td>
<td>0.22</td>
<td>0.098</td>
</tr>
<tr>
<td>3</td>
<td>2.93</td>
<td>0.49</td>
<td>0.167</td>
</tr>
<tr>
<td>4</td>
<td>2.52</td>
<td>0.30</td>
<td>0.119</td>
</tr>
<tr>
<td>5</td>
<td>2.17</td>
<td>0.21</td>
<td>0.097</td>
</tr>
<tr>
<td>6</td>
<td>2.34</td>
<td>0.33</td>
<td>0.141</td>
</tr>
<tr>
<td>7</td>
<td>2.33</td>
<td>0.12</td>
<td>0.052</td>
</tr>
<tr>
<td>8</td>
<td>2.04</td>
<td>0.15</td>
<td>0.074</td>
</tr>
<tr>
<td>9</td>
<td>2.29</td>
<td>0.06</td>
<td>0.026</td>
</tr>
<tr>
<td>10</td>
<td>2.12</td>
<td>0.07</td>
<td>0.033</td>
</tr>
<tr>
<td>11</td>
<td>2.27</td>
<td>0.11</td>
<td>0.048</td>
</tr>
<tr>
<td>All subjects</td>
<td>2.37</td>
<td>0.35</td>
<td>0.148</td>
</tr>
</tbody>
</table>

*Note: Means are based on the average score across the three raters (six trials).*
Generalizability

Variability of MVPA for SOFIT was examined using a participant by rater by trial G-study. True differences in physical activity levels are expected and, therefore; the participant variable was the largest source of variance in MVPA level (94.23%). The second largest source of variance for SOFIT was the residual term which was approximately 3% of total variance indicating that there was little random measurement error. No other facets or interactions had high proportions of variance which indicates that trial, rater, and two-way interaction terms were associated with a relatively low level of measurement error. The index of dependability (phi) for SOFIT was 0.98, indicating there is strong evidence of the ability to generalize the children's observed MVPA scores to their universe score.

Table 3
Variance Component Estimates and their Relative Magnitudes for SOFIT

<table>
<thead>
<tr>
<th>Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>Estimated Variance Components</th>
<th>Relative Magnitude*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants (p)</td>
<td>13634</td>
<td>10</td>
<td>1363.42</td>
<td>224.66</td>
<td>94.23%</td>
</tr>
<tr>
<td>Rater (r)</td>
<td>125.62</td>
<td>2</td>
<td>62.81</td>
<td>1.60</td>
<td>0.67%</td>
</tr>
<tr>
<td>Trial (t)</td>
<td>39.33</td>
<td>1</td>
<td>39.33</td>
<td>0.71</td>
<td>0.30%</td>
</tr>
<tr>
<td>p X r</td>
<td>189.54</td>
<td>20</td>
<td>9.48</td>
<td>0.41</td>
<td>0.17%</td>
</tr>
<tr>
<td>p X t</td>
<td>119.68</td>
<td>10</td>
<td>11.97</td>
<td>1.99</td>
<td>0.83%</td>
</tr>
<tr>
<td>r X t</td>
<td>19.64</td>
<td>1</td>
<td>19.64</td>
<td>1.06</td>
<td>0.44%</td>
</tr>
<tr>
<td>Residual</td>
<td>167.73</td>
<td>21</td>
<td>7.98</td>
<td>7.99</td>
<td>3.3%</td>
</tr>
<tr>
<td>(p X r X t,e)</td>
<td>167.73</td>
<td>21</td>
<td>7.98</td>
<td>7.99</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

Note. * Relative magnitude was calculated using estimated variance divided by the total variance.
Variance estimates of activity scores (mean interval scores) were computed for CARS. As expected the largest variability was due to person with a variance of 49.65%. For CARS, the highest sources of variance after participants were the rater facet (31.49%) and the participant-by-rater interaction term (15.41%). Since terms including rater account for a high proportion of variance there is indication that there were discrepancies between the raters’ coding of participants’ physical activity levels. Contributions of the trial and other interaction terms to the total variance were negligible. This indicates that trial and residual error do not contribute significantly to the total variance. With the testing protocol used (3 raters X 2 trials) CARS had a phi of 0.75. This provides moderate evidence that the children’s observed activity scores may be generalized to their universe score.

Table 4
Variance Component Estimates and their Relative Magnitude for CARS

<table>
<thead>
<tr>
<th>Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>Estimated Variance Components</th>
<th>Relative Magnitude**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants (p)</td>
<td>4.781</td>
<td>10</td>
<td>0.478</td>
<td>0.0721</td>
<td>49.65%</td>
</tr>
<tr>
<td>Rater (r)</td>
<td>2.114</td>
<td>2</td>
<td>1.057</td>
<td>0.0457</td>
<td>31.49%</td>
</tr>
<tr>
<td>Trial (t)</td>
<td>0.002</td>
<td>1</td>
<td>0.002</td>
<td>0*</td>
<td>0%</td>
</tr>
<tr>
<td>p X r</td>
<td>0.983</td>
<td>20</td>
<td>0.049</td>
<td>0.0224</td>
<td>15.41%</td>
</tr>
<tr>
<td>p X t</td>
<td>0.025</td>
<td>10</td>
<td>0.003</td>
<td>0*</td>
<td>0%</td>
</tr>
<tr>
<td>r X t</td>
<td>0.007</td>
<td>1</td>
<td>0.007</td>
<td>0.0002</td>
<td>1.46%</td>
</tr>
<tr>
<td>Residual</td>
<td>0.101</td>
<td>21</td>
<td>0.005</td>
<td>0.0048</td>
<td>3.31%</td>
</tr>
<tr>
<td>(p X r X t,e)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. *Negative variance components were set to zero. ** Relative magnitude was calculated using estimated variance divided by the total variance.
Validity evidence was determined in three ways for each instrument using Pearson product moment correlations. First correlations for each subject were calculated separately. The correlation between the mean score given by the raters for each interval and the interval accelerometer count was calculated (see Table 5). Correlations for SOFIT were low and ranged from 0.44 to +0.39 (Median=0.08). Correlations for CARS ranged from -0.49 to +0.79 (Median=0.38). Negative correlations indicate that there was an inverse relationship between the accelerometer data and the systematic observation data. Negative correlations would indicate that either the accelerometers or systematic observation instruments are measuring behavior inaccurately by either under- or over-estimating physical activity levels.

<table>
<thead>
<tr>
<th>Subject</th>
<th>SOFIT</th>
<th>CARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.04</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>0.31</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>0.39</td>
<td>0.56</td>
</tr>
<tr>
<td>4</td>
<td>0.17</td>
<td>0.43</td>
</tr>
<tr>
<td>5</td>
<td>-0.44</td>
<td>-0.52</td>
</tr>
<tr>
<td>6</td>
<td>0.28</td>
<td>0.76</td>
</tr>
<tr>
<td>7</td>
<td>-0.01</td>
<td>0.43</td>
</tr>
<tr>
<td>8</td>
<td>0.1</td>
<td>0.17</td>
</tr>
<tr>
<td>9</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>10</td>
<td>0.08</td>
<td>0.37</td>
</tr>
<tr>
<td>11</td>
<td>-0.26</td>
<td>-0.10</td>
</tr>
</tbody>
</table>
The second correlation method correlated the overall activity score for all subjects with the total accelerometer counts for all subjects. For SOFIT, the overall activity score used was percent MVPA. This resulted in an extremely low correlation of 0.05. For CARS, the overall activity score used was a mean of the interval scores. This resulted in a low correlation of 0.44.

Finally, a correlation was calculated using every interval score and every interval accelerometer count for all subjects. The correlations were 0.10 and 0.61 for SOFIT and CARS, respectively.

DISCUSSION

The results of the present study provide evidence that using 3 raters and 2 trials SOFIT has high reliability (phi=0.98). This provides strong evidence that SOFIT coding scores may be generalized to the individual's true universe score. Error variance due to trial, rater, and the trial by rater interaction were all low (0.30%, 0.67%, and 0.44%), which indicates that different raters and trials are not sources of error for SOFIT. These small variance components indicate that the training provided for SOFIT in this study was adequate to code physical activity patterns of children with MR. No previous studies have used G-theory to investigate SOFIT's psychometric properties, however; the low error associated with raters lends weight to previous studies which suggested SOFIT had high interrater reliability (McKenzie et al., 1991).
As a follow up calculation, indices of dependability were determined using a decision study (d-study) to determine the optimal measurement protocol (see Table 6). SOFIT maintained a high level of reliability (\(\phi=0.94\)) with only one rater and trial.

<table>
<thead>
<tr>
<th>Number of Raters</th>
<th>Number of Trials</th>
<th>Phi coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.96</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.94</td>
</tr>
</tbody>
</table>

The results of the present study provide evidence that CARS has moderately high dependability (\(\phi=0.75\)). This provides moderate evidence that CARS scores can be generalized to true universe scores. Error variance due to rater (31.49%) and the participant-by-rater interaction (15.41%) accounted for the largest proportions of variance other than participants which indicates the rater facet may need to be adjusted to achieve a higher level of generalizability. Raters received seven hours of training on the two instruments with inter-rater agreement of 87% to 93%. Puhl et al. (1990) provided raters with 30 hours of training with CARS over an 8-week period prior to data collection and interrater reliability estimates. The results of the G-theory analysis indicate that further training is necessary. The low variance associated with trials and the interaction terms including trials indicates that trial does not contribute as a source of error in this study. No previous studies
have been conducted using G-theory with CARS but the results of this study lend
weight to the reliability of the instrument cited in Puhl et al. (1990).

Indices of dependability were also calculated for CARS using a decision
study (d-study) to determine the optimal measurement protocol (see Table 7). With
the level of training provided in this study CARS would need four raters with one
trial to reach a dependability level of 0.80, which is a commonly cited minimum
acceptance level in the literature (Shavelson & Webb, 1991). Using only one rater
and trial would not be a sound decision with the low index of dependability
($\phi = 0.50$). Even with one million trials a single rater would have an index of
dependability of 0.50. Given the low indices of dependability with one rater CARS
would require a more extensive training protocol before being implemented with a
single rater.

<table>
<thead>
<tr>
<th>Number of Raters</th>
<th>Number of Trials</th>
<th>Phi Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.80</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The validity evidence for SOFIT was not strong. Using within-subject
correlations between interval scores and interval accelerometer count the within
subject correlations ranged from -0.44 to +0.36 (Median=0.08). CARS revealed
similar values using mean interval score and interval accelerometer count. The
within subject correlations ranged from -0.46 to +0.79 (Median=0.38). Negative
correlations are a concern because they indicate a tendency for SOFIT and CARS scores to be lower when activity levels are higher or vice versa. This trend in the data opposes the evidence provided in Hodge's study in which within subject correlations ranged between 0.06 and 0.90. The results in the present study may differ from Hodge's because Hodge used heart rate monitors and the present study used accelerometers. Also, the present study synchronized data collection intervals and Hodge's study did not.

Using a more conventional method of correlation a single coefficient was calculated for SOFIT using the MVPA and the total accelerometer count for each subject ($r = 0.05$). Also, the correlation between all SOFIT ratings and interval accelerometer counts was determined to be 0.10. These correlations indicated there is almost no association between the MVPA and accelerometer counts. No previous literature cited correlation coefficients for SOFIT, however; this result differs from by Rowe et al. (1997) who provided evidence for validity of SOFIT levels citing significant differences in heart rates between SOFIT levels. Results of the present study and the study conducted by Rowe et al. (1997) may differ because the studies were conducted in different settings. Rowe's study was conducted by having the participants spend five continuous minutes at each SOFIT level and collecting heart rate data. Horvat and Franklin (2001) suggested when measuring physical activity of individuals with MR "physical activity must be measured in the natural environment to obtain an accurate assessment" (p.189). The present study was conducted in an authentic activity environment.
CARS provided somewhat stronger evidence for validity than SOFIT. A correlation between the mean interval scores for each subject and the total accelerometer count for each subject resulted in a correlation of 0.44. Also, a correlation was calculated using each mean interval score and interval accelerometer count for all subjects \( (r = 0.61) \). Previously cited validity evidence did not use correlations. Validity evidence was reported in Puhl et al. (1990) citing significant differences in VO2 uptake at the different CARS levels. As with the Rowe et al. (1997) study, this testing took place in a laboratory setting rather than an authentic activity setting. This may contribute to the differences in the findings.

Two potential explanations for the low validity coefficients will be raised in the following discussion. The first explanation is that accelerometers with hand placement may be an inappropriate criterion measure for physical activity levels. The second explanation is individuals with MR are more sedentary than their peers without MR and SOFIT and CARS are unable to differentiate between the fitness levels of children with MR.

The first potential explanation for the low validity coefficients is that the accelerometers may have been an inappropriate criterion measure for physical activity. The lessons in which activity was recorded were sometimes sedentary in nature with a great deal of arm movement (e.g. rhythmic gymnastic ribbons, parachute). This may have resulted in a high activity count on the accelerometer because the accelerometer was placed on the child’s arm. It is possible that a child may have been seated shaking a parachute, which would be a Level 2 activity in
both CARS and SOFIT, but their accelerometer would register a great deal of movement because their arm was constantly moving. An overestimation of physical activity levels by the accelerometer would provide an explanation for the negative correlations found within subjects. A correlation between the MVPA for each child from SOFIT and the mean interval score from CARS was calculated to see if there was a positive association between the physical activity level ratings from each instrument. The correlation was 0.76 which indicates that when SOFIT rates a person at a higher activity level CARS has a tendency to also rate that person at a high activity level, and vice versa. This high moderate correlation provides some evidence that accelerometers may have been a source of error.

The second explanation for the low validity coefficients stems from the sedentary behavior of the individuals with MR (Deener & Horvat, 1995). As was addressed in the introduction, it was uncertain whether systematic observation tools could effectively measure physical activity levels of individuals with MR. The data indicate that the participants in this study did have low levels of physical activity. CARS has an overall mean activity level of the children was 2.37 (SD=0.35) and SOFIT had an average MVPA of 26.47% (SD=14.77%). CARS, in particular, demonstrates the homogeneous nature of the data. The inability of the scales to differentiate between the lower fitness levels leads to a reduction in the variability and can deflate correlations. One noticeable difficulty with SOFIT is that it does not differentiate between different walking speeds so when a child is walking at a
very slow pace he/she is coded at a Level 4, but if another child is walking very quickly they are also coded at a Level 4.

The results of this study indicate that SOFIT and CARS can reliably generalize activity levels from a observed score to a true universe score for individuals with MR. However, the results of this study also illuminate concerns about the validity evidence of systematic observation tools in a natural activity setting. While previous studies have demonstrated that SOFIT and CARS provide good validity evidence, these studies have all been in artificial activity environments. Based on the findings in this study validity evidence for SOFIT and CARS may not generalize to individuals with MR. It appears that SOFIT and CARS cannot discriminate between the physical activity levels of the participants in this study.

A conclusion about which systematic observation tool is more appropriate for use with children with MR is not possible. While both SOFIT and CARS demonstrate good reliability, the validity findings warrant further investigation before a decision could be made between the two tools. Suggestions for future research would include conducting a similar study with a more heterogeneous sample if possible, and using alternative criterion measures such as heart rate monitors.
REFERENCES


CHAPTER 3: SUMMARY

The following summary will discuss each research question presented in the introduction, and suggested future research directions.

RESEARCH CONCLUSIONS

1) How much of the error variance in activity level scores is due to the rater?

The error variance in activity level scores due to the rater was very low for SOFIT (0.67%). This indicated that the training provided to the raters on the SOFIT instrument was adequate for raters to use this tool with individuals with MR.

The error variance in activity level scores due to the rater was higher for CARS (31.49%). This finding indicated that the training provided for raters on the CARS instrument was insufficient. More training would be necessary to minimize the error variance associated with the rater facet.

2) How much of the error variance in activity level scores is due to trial?

The error variance in activity level scores due to trial accounted for 0.30% of the error variance for SOFIT. This number is very low and indicates that there are minimal errors associated with the trial facet.

For CARS, the error variance in physical activity level scores due to trial was a negative number which was changed to zero because it is the estimate of the
actual variance component (Morrow, 1989). An error variance of zero indicates that trial was not a source of error in the physical activity scores.

3) Do systematic observation techniques have high interrater reliability?

Interrater reliability coefficients were not calculated using traditional correlation techniques. G-theory results provide an indication of the reliability of different facets based on the error variance estimates. SOFIT had low error associated with rater (0.67%), rater-by-person (0.17%), and rater-by-trial (0.44%), which indicates that rater is not a source of error in the activity level scores. These low error variances lend support to SOFIT demonstrating a high level of interrater reliability.

Rater (31.49%), and rater-by-person (15.41%), accounted for a substantial amount of error variance in the CARS data. This indicates that there may be some interrater reliability problems and that more training is needed on the tool.

4) Do systematic observation techniques have high intrarater reliability?

The low error variance associated with trial (SOFIT = 0.30%, CARS = 0) and trial-by-person interaction (SOFIT= 0.44%, CARS= 1.46%) in both SOFIT and CARS provides evidence for good intrarater reliability.

5) Do SOFIT and CARS demonstrate validity evidence?

SOFIT did not provide strong evidence for validity. Within subject correlations of SOFIT scores and accelerometer counts ranged from -0.44 to +0.39. Correlations between the MVPA level and total accelerometer count of subjects
was 0.05, and a correlation between all interval scores and all interval accelerometer counts was 0.10.

CARS had slightly higher validity coefficients. Within subject correlations between CARS interval scores and interval accelerometer counts lead to a range of correlations from -0.52 to +0.79. The correlation between overall mean CARS score and total accelerometer counts for each subject was 0.44. The correlation between all interval scores and all interval accelerometer counts was 0.61.

Although not at the commonly accepted level 0.80, CARS demonstrated that it has higher validity estimates than SOFIT based on coding of the same children.

6) What is the optimal measurement protocol for SOFIT and CARS?

The optimal measurement protocol for CARS and SOFIT was determined using d-study. The optimal measurement protocol for SOFIT would be using one rater and one trial. This maintains a phi-coefficient of 0.94. The optimal measurement protocol for CARS with the level of training used in this study would be using 4 raters with 1 trial. This is the least number of raters and trials possible to meet the minimum standard of phi=0.80.

FUTURE RESEARCH DIRECTIONS

Suggestions for future research would begin with conducting a similar study using alternative criterion measures such as heart rate monitors, or accelerometer with alternative placement (e.g. hip placement) to determine if the accelerometers
with wrist placement were a poor criterion measure in this study. Once an appropriate criterion measure of physical activity has been established for individuals with MR it would be recommended that a study be conducted with a more heterogeneous sample of individuals with MR to determine if stronger evidence for validity would be obtained when the data have greater range and variability. Also, as McKenzie (1991) suggests several systematic observation tools should continue to be examined to determine the appropriateness of each for individuals with and without disabilities.

Clearly there is a lack of information about the appropriateness of systematic observation tools for use with individuals with disabilities. Similar research studies investigating the generalizability of systematic observation tools for other groups of individuals with disabilities would be beneficial. Another future direction for systematic research would be to develop a method of applying systematic observation of physical activity to individuals who use wheelchairs. If information regarding the appropriateness of applying systematic observation techniques to individuals with and without disabilities it would provide physical educators or fitness leaders with information to develop curricula and increase the health and fitness of targeted individuals.


APPENDICES
The purpose of this review of the literature is to provide an overview of published work in the areas of physical activity for persons with mental retardation, methods for measuring physical activity, and generalizability theory. The background material will provide a rationale for the current study.

Mental Retardation

Mental retardation (MR) is an intellectual disability defined by three characteristics: a "significantly subaverage intellectual functioning, existing concurrently with related limitations in two or more... adaptive skill areas", which is manifested in the developmental period (American Association on Mental Retardation [AAMR], n.d.). Significant subaverage functioning is considered to be an intelligence quotient (IQ) two or more standard deviations below the mean or an IQ of less than 70. Adaptive behavior is a construct that refers to the extent to which an individual meets the standards expected of an individual their age, gender, and culture, for personal skill development, independence, and social responsibility (Bruininks, Thurlow, & Gilman, 1987). In order to be diagnosed with MR individuals must have deficits in two or more adaptive skill areas. The adaptive skills areas identified by AAMR (n.d.) are communication, home living, community use, health and safety, leisure, self-care, social skills, self-direction,
functional academic, and work. The final criterion for diagnosis is that the
disability must be manifested in the developmental period or prior to the age of 18
years.

Physical Activity and Physical Fitness

The impact of physical activity on the reduction of health risks has been
heavily research in recent years. Government initiatives like Healthy People 2010
have been established to raise awareness of healthy behaviors and set goals for the
nation’s health. Healthy People 2010 named physical activity as one of the ten
leading health indictors which “illuminate individual behaviors, physical and social
environmental factors, and important health system issues that greatly affect the
health of individuals and communities” (United States Department of Health and
Human Services, [USDHHS], 2000). In 1996, the Surgeon General issued a report
outlining the benefits of physical activity and guidelines for physical activity
participation (USDHHS, 1996).

The terms physical fitness and physical activity are often confused and
should not be used synonymously. Physical fitness is “the ability to be physically
active on a regular basis” (Donatelle, Snow, & Wilcox, 1999, p.33). Physical
fitness is divided into two components, health-related physical fitness and skill-
related fitness. Health-related fitness is the component of physical fitness focused
on factors that lead to healthy lifestyle and prevention of disease and is commonly
divided into five components; (a) aerobic fitness, (b) flexibility, (c) muscular strength, (d) muscular endurance, and (e) body composition (Pangrazi, 2001). Skill-related fitness is the component of fitness that includes (a) static and dynamic balance, (b) power, (c) agility, (d) speed, and (e) coordination. These perceptual-motor skills require processing abilities and play an important role in everyday activities for all people (Pangrazi, 2001). Physical activity is defined as body movement, produced by the skeletal muscles, which results in energy expenditure. Physical activity includes exercise and activities of daily living like walking upstairs, performing household chores, and carrying groceries (USDHHS, 1996).

Recently, the focus for educating the public about health and wellness has shifted from physical fitness to physical activity (Canadian Society for Exercise Physiology [CSEP], 1999). This shift has educated individuals about the importance of daily physical activity rather than requiring stringent exercise guidelines and vigorous activity. This shift in emphasis is based in research that indicates moderate intensity activities performed in activities of daily living can improve health (CSEP, 1999). The Surgeon General recommends that people of all ages accumulate 30 minutes of daily physical activity at a moderate level of intensity in order to achieve health benefits (USDHHS, 1996). Encouraging moderate intensity lifestyle activities like dancing and gardening give health professionals a broader and gentler approach to physical activity prescription which should allow individuals to find pleasant, useful, and satisfying options for
being active. With this emphasis being placed on physical activity a lifetime fitness approach to better personal health is increasingly common.

Personal health has been conceptualized to have six dimensions: physical health, social health, emotional health, environmental health, spiritual health, and intellectual health (Donatelle, Davis, Munroe, & Munroe, 1998). Participating in regular physical activity can benefit multiple dimensions of personal health.

Physical health refers to the dimension of health that “includes body shape and size, sensory acuity, susceptibility to disease and disorders, body functioning, and recuperative ability” (Donatelle et al, 1998, pg.3). Physical health is the dimension of health most significantly affected by regular participation in physical activity.

Paffenbarger and colleagues have studied the long-term effects of physical activity on physical health in several studies using a large sample of Harvard University and University of Pennsylvania alumni (e.g. Paffenbarger, Wing, and Hyde, 1978). The alumni completed a self-report questionnaire regarding their level of participation in physical activity (based on stairs climbed, blocks walked, and sports played), doctor diagnosed diseases, tobacco smoking, and parental disease or death. Helmrich, Ragland, and Paffenbarger (1994) studied 5990 male Harvard alumni and found that incidence rates of adult onset diabetes declined as level of energy expenditure increased. A study by Manson et al. (1991) supports this conclusion. In their eight year follow-up study of 87,253 women, participation in vigorous activity was associated with a reduction in cases of adult onset diabetes.
Physical activity reduces the risk of heart disease which is the leading cause of death in the United States (American Heart Association, n.d.). The study by Paffenbarger et al. (1978) of 16,936 male Harvard Alumni indicated that men with lower activity levels were at a higher risk for first heart attack than their more active peers. Sesso, Paffenbarger, and Lee (2000) conducted a 16-year follow-up study of 12,516 Harvard alumni examining the effect of physical activity on coronary heart disease. The data suggest strong associations between lower risk of coronary heart disease and higher levels of physical activity. Using a sample of 1,564 female University of Pennsylvania alumni Sesso, Paffenbarger, Ha, and Lee (1999) found that amount of walking was inversely related to a reduction in the risk of cardiovascular disease.

Reducing the risk of obesity is another benefit of participation in physical activity. Ching et al. (1996) examined the relationships between non-sedentary activities, sedentary activities, and risk for being overweight in a two year follow-up study in a sample of 17,795 men. The results indicated that a higher level of participation in non-sedentary activity was related to a reduced risk of becoming overweight. Higher levels of sedentary activities were independently associated with risk of becoming overweight.

Dalsky et al. (1988) conducted an intervention study with sedentary women to investigate the influence of weight bearing physical activity (walking, jogging, and stair-climbing) on bone mineral density. Results indicated there was a significant increase in bone mineral density for the exercise group after nine
months of exercise. This increase was maintained through the 22 months of continued exercise. There was no change in bone density in the control group.

Low bone density, a risk factor in osteoporosis, can cause bone to be fragile and make individuals more susceptible to fractures.

The relationship between physical activity and cancer has inconsistent findings. According to the Surgeon General’s report participation in physical activity may be associated with a reduced risk for colon cancer and breast cancer later in life (USDHHS, 1996). More research in this area is needed.

Physical activity has a significant impact on all cause mortality. Blair et al. (1995) conducted a longitudinal study with 9,777 men. Two clinical examinations for physical fitness levels and risk for all-cause mortality were performed with a one to 18 year break between testing sessions. Men who were unfit at both visits had the highest death rates, men who were fit at both visits had the lowest death rates, and men who had changed fitness levels had intermediate death rates. Men who back fit between the two testing sessions reduced their risk of all-cause mortality by 44%. In a similar study with males and females Blair et al. (1989) found that all-cause mortality rates were inversely related to physical activity levels in both males and females.

Participation in regular physical activity is also beneficial for improving psychological health. While not studied as extensively as physical health some studies have addressed the impact of physical activity on depression, well-being, anxiety, and self-efficacy (Brosse, Sheets, Lett, & Blumenthal, 2002). Brosse et al.
indicated that regular physical activity had a positive impact on depression.

Hassmen, Koivula, & Uuetla (2000) provide a study which indicates individuals who participate in regular physical activity experience less depression, anger, cynical distrust, and stress than individuals who do not regularly participate in physical activity. Those who were more active also felt more in control of their health, more socially intergrated, and perceived themselves as both healthier and fitter.

Because of the significant number of conditions resulting from sedentary lifestyles the Surgeon General has labeled physical inactivity a major public health problem in the United States (US). The 1996 Behavioral Risk Factor Surveillance System (BRFSS), a national survey of approximately 120,000 randomly selected adults representing all 50 states, indicated that 43.1% of adults participate in some physical activity but not enough to achieve health benefits and 29.2% are inactive. Only 28% of adults participate in physical activity at the recommended level.

Participation in regular physical activity declines with age (CDC, 1996).

Similar studies have been designed to measure current physical activity levels of children and adolescents. According to Healthy People 2010 approximately 65% of adolescents participated in the recommended amount of vigorous physical activity in 1999 (USDHHS, 2000). A study by Pratt, Macera, and Blanton (1999) indicated there is a tendency for participation in physical activity to decline with age in both genders. Because of the decline in physical activity, and the risks associated with sedentary lifestyles it is very important to
ensure children learn the importance of daily physical activity and physical activity levels are improved upon.

Physical Activity, Physical Fitness, and Persons with Mental Retardation

Physical activity is very important for adults with MR as they often have more leisure time than individuals without MR (Eichstaedt & Lavay, 1992). Individuals with MR also have more sedentary lifestyles than individuals without MR (Deener & Horvat, 1995; Fernhall, 1993; Fernhall et al., 1996). Because physical activity levels decrease with age it is important to encourage participation in physical activity to children with MR so that it may become a lifetime behavior (Deener & Horvat, 1995). Benefits of physical activity for individuals with MR include improvements in social interaction, job productivity, worthy use of leisure time, and resistance from disease (Eichstaedt & Lavay, 1992).

The association between physical activity and physical fitness has been demonstrated in the literature (Pate et al., 1990). In their study of 2,352 children Pate et al. (1990) found significant relationships between physical activity and physical fitness measures. Individuals who participated in regular physical activity were more likely to have higher physical fitness ratings. This is of particular importance when trying to increase physical activity levels in children with MR because disparities in physical fitness exist between individuals with and without MR.
Statistics from the “Healthy People 2000 Review” (1999) indicate there are disparities between the health and fitness levels of individuals with and without disabilities. Individuals with disabilities reported greater incidence of being overweight and adverse health effects due to stress. Although there was a noticeable decline in individuals with disabilities classified as inactive between 1985 and 1995 (35 percent to 29 percent) the target level of 20 percent set for the year 2000 has not yet been met.

One of the target goals for Healthy People 2010 is to “promote the health of people with disabilities, prevent secondary conditions, and eliminate disparities between people with and without disabilities in the U.S. population” (USDHHS, 1999, p.6-6). Some secondary conditions, like osteoporosis, can be avoided by participating in resistance training. Encouraging physical activity in individuals with disabilities can reduce risk for secondary conditions and help eliminate disparities.

Research into health and fitness levels of individuals with MR has demonstrated there are health and fitness disparities between individuals with and without MR. Individuals with mental retardation are at a greater risk for poor cardiovascular health due to poor physical activity habits, high body fat, and sedentary lifestyle (Deener & Horvat, 1995; Fernhall, 1993; Fernhall et al, 1996). Cardiovascular endurance is considered the best single indicator of overall physical fitness (American College of Sports Medicine, 1990). The average deficit in
cardiovascular fitness across several studies is 25 to 30% when compared with expected values for individuals without disabilities (Fernhall, 1993).

Body composition is also a poor area for individuals with MR. Individuals with MR have a greater prevalence of obesity than individuals without MR (Fernhall, 1993). Pitetti, Yarmer, and Fernhall (2001) found that children with MR tended to have higher body mass index (BMI) than children without MR.

The study by Reid, Montgomery, and Seidl (1985) indicated that males and females with MR fell into the 1-2 and 4 percentile respectively for push-ups. They also performed very poorly in sit-ups. These results indicate that individuals with MR have poor muscular endurance. Nordgren (1970) studied upper and lower body isometric strength and found that individuals with MR performed at 71% and 78% of their expected values when compared with individuals without MR.

Measuring Physical Activity

Recent research has used many techniques to attempt to quantify physical activity levels in both children and adults. Self-report methods, heart rate monitors, pedometers, accelerometers, the doubly labeled water technique, and systematic observation have all been used (Rowe, Schuldheizs, & van der Mars, 1997).

Heart rate monitors are a commonly used biological marker to measure physical activity levels. Strath et al. (2000) correlated heart rates with rate of oxygen uptake (VO2) scores for a 15 minute period of physical activity. Heart rate
monitors were programmed to record heart rates at specified intervals and were correlated with VO2 scores at the same time interval. Results of this study indicated that heart rates may be used to accurately quantify physical activity. Heart rate monitors can also provide information about energy expenditure and intensity, and are inexpensive compared to accelerometers and laboratory based tests (Dishman, Washburn, & Schoeller, 2001).

Welk et al. (2000) conducted a study of the utility of a pedometer to assess physical activity levels in adults. Energy expenditure levels based on self-report methods were compared to daily pedometer readings. Although there was a trend that lower number of steps taken in a day correlated with lower energy expenditure there are two drawbacks to the results of this study. One drawback is that the number of steps taken is recorded in one interval for the entire day so there is no indication of how the steps are spread over time. Greater gains in health take place when physical activity is at a moderate level for an extended period of time so having a measure that considers the intensity is important. The second drawback is that if a person is running they will actually take fewer steps while having higher energy expenditure than a person who walks for the same period of time.

Accelerometers are small devices used to measure acceleration of the body or limbs. Typically worn on the wrist, ankle, or hip, accelerometers can provide information about frequency, duration, and intensity of movement because they collect and store data sequentially over time (Dishman et al., 2001). Nichols et al. (1999) looked at the validity, reliability, and calibration of accelerometers. The
results of this study indicated that accelerometers are reliable over time and are sensitive to changes in speed. Based on these results accelerometers appear to be able to distinguish between low, moderate, and high levels of physical activity, and may be used to estimate energy expenditure.

Accelerometers provide data in the form of activity counts. The activity count represents the summed magnitude of the accelerations. This gives an indication of how much the individual moved in the interval recorded. In the study by Nichols et al. (1999), 90% of the variance in energy expenditure was explained by the accelerometer activity counts.

Systematic Observation

Observational techniques like eye-balling and check-lists have a tendency toward bias and do not provide specific information for the individual being observed. Systematic observation can be more reliable, objective, and specific than other observational techniques. Systematic observation is the observation and recording of a behavior based on specifically designed guidelines and coding procedures (van der Mars, 1989).

Systematic observation tools can be used to record information pertaining to behaviors, events, and episodes. Only observable behavior can be measured. Inferences to intent, internal states, and private events are beyond the scope of
systematic observation. Observations may be made based on form of skills, or outcomes and consequences of behavior (van der Mars, 1989).

Continuous or duration recording and interval recording are tactics used to code (record) behaviors with systematic observation tools (Test & Heward, 1984). Duration recording is used to determine the amount of time an individual spends performing a specific behavior. An individual coding with this method would use a stopwatch to start every time the behavior is displayed and would stop when the behavior stops. The watch would be restarted and stopped whenever the behavior took place so that a total elapsed time could be determined. A percentage would be calculated by dividing the total seconds the behavior was displayed by the total seconds observed.

Interval recording may be used as whole-interval sampling, partial-interval sampling, or momentary time sampling. In whole-interval recording the behavior must take place for the entire interval to be recorded. In partial-interval recording the behavior must be present in any part of the interval to be recorded. And in momentary time-sampling the behavior that is taking place at the end of the interval is the one recorded (Test & Heward, 1984). The recording technique selected will depend on what information is to be obtained from the results, and the duration and repeatability of the behaviors to be coded (van der Mars, 1989).

There are many advantages to using systematic observation including that observation can take place in a natural activity setting rather than a research lab, it requires little equipment, and it inexpensive (Epstein et al., 1984, Puhl et al., 1990).
Because it is a direct measure of behavior systematic observation is sometimes used as a criterion measure to validate other methods of measuring physical activity.

Difficulties with systematic observation include potential for participant reactivity, and its limitation to observable behaviors, and descriptive information (van der Mars, 1989). In order to reduce participant reactivity, participants changing behavior because they are being observed, participants should be told an observer and equipment will be present in advance, the observer should be present at the beginning of the class to avoid additional disruption, the observer should be courteous but not interact with the participants, and equipment should always be placed in the same location.

Systematic Observation of Physical Activity

Information obtained through systematic observation may include the type, intensity, duration, physical environment, and social environment of activity (O'Hara et al., 1989). Many systematic observation tools have been developed to measure levels of physical activity (McKenzie et al., 1991; O'Hara et al., 1989; Puhl et al., 1990).

The Children's Activity Rating Scale (CARS) was developed by Puhl et al. (1990) to categorize activities of different intensities, and estimate energy expenditure levels. Five categories were selected and defined based on energy
expenditure in children. These levels ranged from no movement to very fast translocation. Partial-interval recording of behaviors takes place every minute. If a behavior takes place for three seconds then it is coded for that minute. Each level that is performed for at least three seconds is coded. An activity rating score is calculated for that minute using the number, and duration of the behaviors exhibited.

Puhl et al. (1990) reported interrater reliability of CARS using mean percent agreement. Percent agreement between raters for each interval was calculated using the number of physical activity levels coded in agreement, divided by the total number of levels coded for that interval. For example, if rater one coded levels 1, 2, and 3 for an interval and rater two coded levels 2, and 3 for the same interval the percent agreement would be 0.66. Using more than 389 paired observations Puhl et al. (1990) reported interrater reliability of .84.

Validity of the CARS levels was investigated using VO2 and heart rates. Significant differences in the VO2 values for the five CARS levels demonstrate activity levels are differentiated using the CARS system. F-values and p-values were not reported in this study. No psychometric properties have been established for CARS with individuals with mental retardation.

The System for Observing Fitness Instruction Time (SOFIT) was developed by McKenzie et al. (1991). SOFIT is a momentary time sampling and interval recording systematic observation tool that divides physical activity behaviors into five categories ranging from lying down to standing, very active. Levels 4 and 5
are equivalent with moderate to vigorous physical activity (MVPA) (Rowe et al., 1997). An activity score can be computed using the percent of the intervals in which the individual was performing MVPA. Behaviors are coded according to what the participant is doing at the moment the interval ends.

Psychometric properties for SOFIT have been examined with individuals without disabilities. Rowe et al. (1997) conducted a study using heart rates as the dependant variable and the activity levels as the independent variable. Subjects spent five continuous minutes performed each of the SOFIT levels in addition to push-ups and sit-ups.

Internal consistency reliability of heart rates across tasks was calculated using the heart rates for the last four minutes of these intervals. A one-way ANOVA was used to calculate intraclass correlations which ranged from 0.98 for push-ups to 0.99 for all other activities. Standard error of the measurement (SEM) was calculated to be one beat per minute (bpm) for all activities except push-ups (SEM = 1.7 bpm).

Stability reliability of heart rates for each task was calculated using intraclass correlations. The time lag between the two measurement occasions varied from one day to one week. A one-way repeated measures 2 X 4 X 7 (Gender X Age X Task) ANOVA was employed to calculate reliability coefficients. Reliability coefficients ranging from R=0.80 for jogging to R=0.91 for curl-ups. SEM values were also calculated and ranged from 4.1 bpm for curl-ups to 5.7 bpm for jogging.
Intraclass correlations for heart rates of individuals measured on only one day were calculated using a one-way 2 X 4 X 7 (Gender X Age X Task). Reliability coefficients ranged from 0.66 for jogging to 0.84 for curl-ups. SEM values ranged from 5bpm for curl-ups to 7.3 bpm for standing.

McKenzie et al. (1991) examined interrater reliability of SOFIT using an interval by interval calculation. The calculation was performed by taking the number of intervals agreed upon and dividing it by the total number of intervals coded. Based on 2063 intervals two observers had a reliability level after training of 0.88 for the student activity portion of SOFIT (McKenzie et al., 1991).

Validity evidence in this study was examined using heart rate values in a one-way repeated measures 2 X 4 X 7 (Gender X Age X Task) ANOVA. There was a significant task-effect (p<0.0001) indicating that there are significant differences in the mean heart rates by task (SOFIT level). Scheffe post-hoc analysis showed significant difference in the mean heart rates for all coding levels of SOFIT except lying down and sitting.

The validity of SOFIT with children with mental retardation has been investigated by Hodge (2001). Hodge cited concurrent validity using correlations between heart rates and SOFIT scores for 8 children with mental retardation between 0.72 and 0.86. Hodge (2001) also cited that SOFIT has face validity for individuals with mental retardation because there are no physical disabilities that would affect posturing (sitting, lying down) and subsequently scoring of this instrument. However, there are some methodological issues with Hodge’s design.
First, the heart rate monitor intervals and SOFIT intervals were not synchronized. Second, no reliability information was cited. Third, Hodge did not use a conventional correlation method. Heart rate values and SOFIT levels for each person were correlated, therefore; the correlations cited were for individuals. This is a descriptive measure and can not be generalized. Further investigation into the validity of SOFIT for individuals with MR should be conducted.

Generalizability Theory

Generalizability theory (g-theory) is a data analysis technique which allows the user to determine how much error variance in the data is accounted for by different variables. "G-theory is an extension of the intraclass reliability model. G-theory is perhaps the most appropriate methodology for estimating test reliability available today" (Morrow, 1989, p.75).

The design of a G-theory depends on the number and type of facets. A facet is a "dimension on which measures are taken" (Morrow, 1989). Facets in a study may be fixed or random. A random facet is randomly representative of the potential universe of that facet. A fixed facet is when all possible levels of that facet are being measured; there are no more levels to which you may generalize. Using G-theory, variance components are calculated for each facet. Variance components indicate how much of the total variance in the measures can be
explained by each facet and interaction term. The residual variance component includes systematic or random error (Shavelson & Webb, 1991).

G-theory may have fully-crossed or nested designs. A crossed design is a design in which all subjects (or objects of measurement) are crossed with all facets. A nested design occurs when some subjects are not measured on all facets. Depending on if facets are random or fixed the design may be random (all random facets) or mixed (both random and fixed) (Morrow, 1989). Nested facets limit generalizability because in some cases not all variance components can be estimated. Similarly, random facets can be generalized more easily than fixed facets because fixed each level of the fixed facet may need to be studied independently (Shavelson & Webb, 1991).

Generalizablity coefficients (g-coefficients) can be calculated to give an indication of reliability. G-coefficients give an indication of how sound the generalization for an individual’s observed score is to their true universe score. G-coefficients are calculated differently depending on what sort of decisions are to be made from the results. If interpretations pertain to the relative standing of participants then a relative decision is made. G-coefficients for relative decisions are calculated using all the interactions terms between the facets and the object of measurement. If decisions are based on the absolute score for participants and relative standing has no bearing then absolute decisions are made. In the case of absolute decisions all variance components except the object of measurement are used in the calculation of an index of dependability. The index of dependability is
the absolute decision equivalent of the G-coefficient in relative decisions (Shavelson & Webb, 1991).

Following a G-theory analysis a decision study (D-study) can be used to develop an optimal measurement protocol for data collection. The optimal measurement protocol can tell a researcher how many of each facet are necessary to maintain high reliability.
APPENDIX B – INFORMED CONSENT DOCUMENT

INFORMED CONSENT DOCUMENT

Project Title: Psychometric properties of systematic observation tools used to measure physical activity in children with mental retardation.

Investigator: Christina Taylor and Joonkoo Yun, Department Exercise & Sport Science

Research Staff: Soyeun Kim and Jason Russell

PURPOSE

The purpose of this research study is to evaluate the validity and reliability of using systematic observation techniques to measure physical activity levels of children with disabilities. The ability to use systematic observation techniques with individuals would provide researchers and practitioners an effective way of measuring activity levels and a base for providing appropriate services. The intended uses of the project are to complete partial requirements for a Master of Science degree and to publish results in a journal. The purpose of this consent form is to give you the information you will need to decide whether or not to you want your child to participate in the study. Please read the form carefully. Your child may ask any questions about the research, what your child will be asked to do, the possible risks and benefits, your child’s rights as a volunteer, and anything else about the research or this form that is not clear. When all of your questions have been answered, you can decide if you want your child to be in this study or not. This process is called “informed consent”. You will be given a copy of this form for your records.

PROCEDURES

If you agree to participate, data collection for your child will take place during one clinic session. While your child will participate in regular clinic session, your child’s physical activity behavior will be videotaped for 15 minutes and he or she will wear an accelerometer.

The following procedures are involved in this study. Upon arriving at the gym activity portion of clinic your child wear accelerometer on their wrist of their non-dominant side. An accelerometer is a small watch-like device (one inch squared) that measures the acceleration of the body during movement. It gives an estimate of how much your child moves during clinic. While your child participates in regular physical activity session, your child will be videotaped for 15 minutes. The videotape of your child will be viewed by members of the research team only. The
research team will use the two observation techniques being evaluated to estimate your child’s level of physical activity. These tools will be compared with the data collected by the accelerometers to see which tool may be more appropriate to use. At the end of the gym activity time the accelerometer will be removed and your child’s participation in this study will be complete.

**BENEFITS and RISKS**

No additional risk is posed to participants in this study. The potential personal benefits that may occur as a result of your child’s participation in this study are that you will receive information about your child’s activity level during clinic time. The researchers anticipate that society may benefit from this study by providing researchers and practitioners a way to measure activity levels and to compare activity levels between individuals with and without disabilities.

**CONFIDENTIALITY**

Records of participation in this research project will be kept confidential to the extent permitted by law. However, federal government regulatory agencies and the Oregon State University Institutional Review Board (a committee that reviews and approves research studies involving human subjects) may inspect and copy records pertaining to this research.

Videotapes of your child’s activity will be labeled by participant number and will not be linked with their personal information. Informed consent forms will be stored in a private location. At the conclusion of the study all documents except the informed consent will be destroyed. In the event of any report or publication from this study, your child’s identity will not be disclosed. Results will be reported in a summarized manner in such a way that your child cannot be identified.

**RESEARCH RELATED INJURY**

In the event of research related injury, compensation and medical treatment is not provided by Oregon State University.

**VOLUNTARY PARTICIPATION**

Taking part in this research study is voluntary. You may choose for your child not to take part at all. If you agree for your child to participate in this study, you may stop their participation at any time. If you decide for your child not to take part, or if you stop your child from participating at any time, your decision will not result in lack of access to Clinic.
QUESTIONS
Questions are encouraged. If you have any questions about this research project, please contact: Christina Taylor, at (541) 737-5927 or by e-mail at taylochr@onid.orst.edu or Dr. JK Yun, at (541)-737-8584 or by e-mail at JK.Yun@orst.edu. If you have questions about your rights as a participant, please contact the Oregon State University Institutional Review Board (IRB) Human Protections Administrator, at (541) 737-3437 or by e-mail at IRB@oregonstate.edu or by mail at 312 Kerr Administration Building, Corvallis, OR 97331-2140.

Your signature indicates that this research study has been explained to you, that your questions have been answered, and that you agree for your child to take part in this study. You will receive a copy of this form.

Participant's Name (printed): ____________________________________________

______________________________________________________________
(Signature of Participant)
(Date)

______________________________________________________________
(Signature of Parent/Guardian or Legally Authorized Representative)
(Date)

RESEARCHER STATEMENT

I have discussed the above points with the participant or, where appropriate, with the participant’s legally authorized representative, using a translator when necessary. It is my opinion that the participant understands the risks, benefits, and procedures involved with participation in this research study.

______________________________________________________________
(Signature of Researcher)
(Date)
APPENDIX C – INSTITUTIONAL REVIEW BOARD APPROVAL

OREGON STATE UNIVERSITY
INSTITUTIONAL REVIEW BOARD
312 Kerr Administration Building, Corvallis, Oregon 97331-2140
E-MAIL: IRB@oregonstate.edu, PHONE: (541) 737-3437, FAX: (541) 737-3093

REPORT OF REVIEW

TO: Joonkoo Yun,  
Exercise and Sport Science

RE: Psychometric Properties of Two Systematic Observation Tools Used to Measure Physical Activity in Children with Mental Retardation (Student Researcher: Christina Taylor)

Protocol No. 2105

The referenced project was reviewed under the guidelines of Oregon State University’s Institutional Review Board (IRB). The IRB has approved the application. This approval will expire on 2/13/2004. This new request was reviewed at the Expedited level. A copy of this information will be provided to the full IRB committee.

Enclosed with this letter please find the original informed consent document for this project, which has received the IRB stamp. The original informed consent document has been stamped to ensure that only current, approved informed consent forms are used to enroll participants in this study. All participants must receive the IRB-stamped informed consent document. Make copies of this original as needed.

* Any proposed change to the approved protocol, informed consent form(s), or testing instrument(s) must be submitted using the MODIFICATION REQUEST FORM. Allow sufficient time for review and approval by the committee before any changes are implemented. Immediate action may be taken where necessary to eliminate apparent hazards to subjects, but this modification to the approved project must be reported immediately to the IRB.

* In the event that a human participant in this study experiences an outcome that is not expected and routine and that results in bodily injury and/or psychological, emotional, or physical harm or stress, it must be reported to the IRB Human Protections Administrator and the ADVERSE EVENT FORM must be completed within three days of the occurrence using the MODIFICATION REQUEST FORM and the ADVERSE EVENT FORM as needed.

Before the expiration date noted above, a Status Report will be sent to either close or renew this project. It is imperative that the Status Report is completed and submitted by the due date indicated or the project must be suspended to be compliant with federal policies.

If you have any questions, please contact the IRB Human Protections Administrator at IRB@oregonstate.edu or by phone at (541) 737-3437.

Dr. Anthony Wibro 
Institutional Review Board Chair

Date: 2/14/03

pc: 2105 file
APPENDIX D – SAS COMMAND FILE FOR G-THEORY ANALYSIS

options ps = 55
options ls = 79
pageno = 1;
data sofit;
input
   r1t1 r1t2 r2t1 r2t2 r3t1 r3t2
   rater=1; trial=1; score=r1t1; output;
   rater=1; trial=2; score=r1t2; output;
   rater=2; trial=1; score=r2t1; output;
   rater=2; trial=2; score=r2t2; output;
   rater=3; trial=1; score=r3t1; output;
   rater=4; trial=1; score=r3t2; output;
[insert data here]
Proc varcom method=type1;
   class sub rater trial
   model score=sub rater trial
      sub*rater
      sub*trial
      rater*trial;
Proc glm;
   class sub rater trial
   model score=sub rater trial
      sub*rater
      sub*trial
      rater*trial;
run;

Data for SOFIT

   r1t1 r1t2 r2t1 r2t2 r3t1 r3t2
   50.9 55 58.6 56.9 44.8 51.7
   32.2 30.9 33.9 25.4 28.8
   32.8 36.6 41.4 40.0 35.4 33.9
   38.9 37.9 32.7 40.4 32.7 35.7
   16.9 13.3 13.6 15.3 13.6 13.6
   45 45 46.6 45 40 40
   31.7 28.8 30 33 30 30
   6.8 6.7 3.6 5.5 5.5 5.5
   10 15 10 10 10 10
13.6 13.3 15.3 16.7 15.3 16.9
13.3 16.6 11.7 31.6 16.6 11.7
Data for CARS

r1t1 r1t2 r2t1 r2t2 r3t1 r3t2

3.02 3.22 2.77 2.58 2.73 2.70
2.53 2.53 2.11 2.1 2.07 2.18
3.54 3.57 2.7 2.67 2.49 2.60
2.87 2.9 2.41 2.49 2.17 2.3
2.4 2.47 2.13 2.2 2.02
2.78 2.75 2.21 2.14 2.07 2.11
2.47 2.47 2.3 2.3 2.27 2.17
2.27 2.17 2.2 1.93 1.87
2.3 2.4 2.23 2.27 2.27 2.27
2.1 2.2 2.13 2.13 2.17 2
2.37 2.4 2.09 2.27 2.27 2.19
APPENDIX E - SOFIT CODING FORM

Coder: ________________
Date: ________________
Subject: ________________
Trial: ______
Tape start time: 6:23:45

| Level 1: Lying Down | 1: 6:24:00 1 2 3 4 5 NA | 31: 6:31:30 1 2 3 4 5 NA | Level 2: Sitting | 2: 6:24:15 1 2 3 4 5 NA | 32: 6:31:45 1 2 3 4 5 NA |
| Level 5: Running/Moderate | 5: 6:25:00 1 2 3 4 5 NA | 35: 6:32:30 1 2 3 4 5 NA |             | 6: 6:25:15 1 2 3 4 5 NA | 36: 6:32:45 1 2 3 4 5 NA |
|                        | 7: 6:25:30 1 2 3 4 5 NA | 37: 6:33:00 1 2 3 4 5 NA |             | 8: 6:25:45 1 2 3 4 5 NA | 38: 6:33:15 1 2 3 4 5 NA |
|                        | 9: 6:26:00 1 2 3 4 5 NA | 39: 6:33:30 1 2 3 4 5 NA |             | 10: 6:26:15 1 2 3 4 5 NA | 40: 6:33:45 1 2 3 4 5 NA |
|                        | 11: 6:26:30 1 2 3 4 5 NA | 41: 6:34:00 1 2 3 4 5 NA |             | 12: 6:26:45 1 2 3 4 5 NA | 42: 6:34:15 1 2 3 4 5 NA |
|                        | 13: 6:27:00 1 2 3 4 5 NA | 43: 6:34:30 1 2 3 4 5 NA |             | 14: 6:27:15 1 2 3 4 5 NA | 44: 6:34:45 1 2 3 4 5 NA |
|                        | 15: 6:27:30 1 2 3 4 5 NA | 45: 6:35:00 1 2 3 4 5 NA |             | 16: 6:27:45 1 2 3 4 5 NA | 46: 6:35:15 1 2 3 4 5 NA |
|                        | 17: 6:28:00 1 2 3 4 5 NA | 47: 6:35:30 1 2 3 4 5 NA |             | 18: 6:28:15 1 2 3 4 5 NA | 48: 6:35:45 1 2 3 4 5 NA |
|                        | 19: 6:28:30 1 2 3 4 5 NA | 49: 6:36:00 1 2 3 4 5 NA |             | 20: 6:28:45 1 2 3 4 5 NA | 50: 6:36:15 1 2 3 4 5 NA |
|                        | 21: 6:29:00 1 2 3 4 5 NA | 51: 6:36:30 1 2 3 4 5 NA |             | 22: 6:29:15 1 2 3 4 5 NA | 52: 6:36:45 1 2 3 4 5 NA |
|                        | 23: 6:29:30 1 2 3 4 5 NA | 53: 6:37:00 1 2 3 4 5 NA |             | 24: 6:29:45 1 2 3 4 5 NA | 54: 6:37:15 1 2 3 4 5 NA |
|                        | 25: 6:30:00 1 2 3 4 5 NA | 55: 6:37:30 1 2 3 4 5 NA |             | 26: 6:30:15 1 2 3 4 5 NA | 56: 6:37:45 1 2 3 4 5 NA |
|                        | 27: 6:30:30 1 2 3 4 5 NA | 57: 6:38:00 1 2 3 4 5 NA |             | 28: 6:30:45 1 2 3 4 5 NA | 58: 6:38:15 1 2 3 4 5 NA |
|                        | 29: 6:31:00 1 2 3 4 5 NA | 59: 6:38:30 1 2 3 4 5 NA |             | 30: 6:31:15 1 2 3 4 5 NA | 60: 6:38:45 1 2 3 4 5 NA |

Mean SOFIT score (mean score for all intervals): ______
% time in MVPA (% of intervals at level 4 or 5): ______
APPENDIX F - CARS CODING FORM

Coder __________
Date __________
Subject# 11
Trial # ____
Tape start time _6:23:45

<table>
<thead>
<tr>
<th>Interval</th>
<th>CARS SCORE</th>
<th>ACTIVITY SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6:23:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>2</td>
<td>6:24:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>3</td>
<td>6:25:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>4</td>
<td>6:26:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>5</td>
<td>6:27:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>6</td>
<td>6:28:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>7</td>
<td>6:29:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>8</td>
<td>6:30:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>9</td>
<td>6:31:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>10</td>
<td>6:32:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>11</td>
<td>6:33:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>12</td>
<td>6:34:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>13</td>
<td>6:35:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>14</td>
<td>6:36:45</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>15</td>
<td>6:37:45</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

Mean activity score (average of scores over all intervals): ______

**LEVEL 1** (Stationary/motionless, resting for more than 3 seconds or more; head, finger or foot movement only)

**LEVEL 2** (Stationary/Movement of the limbs or trunk, very easy, arm trunk and/or leg movements without moving the entire body from one place to another)

**LEVEL 3** (Translocation, slow speed/easy, moving body from one location to another)

- sit ups and push-ups

**LEVEL 4** (Translocation, medium speed/moderate, moving the whole body from one location to another)

- 3 or more forward or backward rolls

**LEVEL 5** (Translocation, fast or very fast, hard, moving the whole body from one location to another)

- push-ups