RELIABILITY OF THE MOVEMENT ASSESSMENT BATTERY FOR CHILDREN – SECOND EDITION: AGE BAND 2

A Thesis Presented to the
School of Kinesiology
Lakehead University

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

Supervisor: Dr. Eryk Przysucha

Committee Members: Dr. Carlos Zerpa & Mr. Vineet Johnson

Date: July 2016

Abstract

The purpose of this study was to examine the reliability of age band 2 of the Movement Assessment Battery for Children – Second Edition (MABC-2), using test re-test, internal consistency, and standard error of measurement (SEM) reliability estimates. To date, only one study has been conducted on the reliability of age band 2 (7 to 10 year olds) (Holm, Tveter, Aulie, & Stuge, 2013), thus more research is warranted as many practitioners consider this tool as a gold standard in this area. Forty typically functioning children (18 boys, 22 girls) between the ages of 7 and 10 (M = 9 years, 0 months, 5 days, SD = 1 year, 0 months, 15 days) participated in this study. Each child completed two, thirty-minute sessions, one to two weeks apart, and was examined by the same researcher in the same laboratory setting. Intra-class correlation (ICC) coefficient was used to examine the test-retest stability of the Total Impairment Score (TIS) and three sub-section scores. Cronbach's alpha was used to examine the internal consistency of the items. Lastly, the SEM was computed to infer the magnitude of absolute reliability for each of the four scores.

The normality, skewness and kurtosis of ICC, Cronbach's alpha and SEM were tested first across standard, component, and percentile scores. The analyses showed that for the standard scores, aiming and catching and balance sessions at time 1, and balance items at time 2 did not meet the assumptions of normality. The component scores showed similar results, while the assumption of normality was jeopardized for six out of eight percentile scores. The analyses of variance and scatterplots supported the inferences emerging from the normality tests. Overall, the normality results revealed that the standard scores best represented a normal distribution and met the necessary statistical assumptions.

The analyses of test-retest reliability, for standard scores, revealed ICC coefficients of .67 for manual dexterity, aiming and catching, and balance sub-sections, and 0.65 for the Total Test

Score (TTS). These results indicated that the coefficients were approaching a moderate degree of reliability. The test-retest reliability of the component scores revealed a similar pattern of results with ICC coefficients of 0.70, 0.71, 0.62, and 0.55 for the TTS, manual dexterity, aiming and catching, and balance sub-components, respectively. Lastly, the percentile scores revealed coefficients that ranged from 0.42 to 0.68, indicating a weak to moderate reliability. Overall, the standard scores were the most homogenous across the four different types of values, however the component scores reported the highest ICC of 0.71. The test-retest reliability revealed here was similar to that evident in the previous study on age band 2 (Holm et al., 2013). However, they were much lower than the coefficients reported in studies conducted on other age bands.

The analysis of the internal consistency revealed poor to questionable reliability among the items for manual dexterity (.61) aiming and catching (.49), and balance (.53) subsections, respectively. Once again, these findings were similar to the study conducted by Holm and colleagues (2013), but considerably lower than the coefficients reported in other literature (Wuang et al., 2012). The analysis of absolute reliability for TTS revealed SEM scores of 1.80, 7.39, and 18.59 for the standard, component, and percentile scores, respectively. Based on these results, it is evident that there is a high degree of measurement error, particularly in the component and percentile scores. The values for the subsections were also the lowest for the standard scores across manual dexterity (2.25), aiming and catching (1.78), and balance (1.92) sections. The SEM for the component scores was higher, however the highest values were revealed for percentile scores across all three subsections.

Overall, the analysis of different aspects of reliability for the TTS and the three subsections, across the different scores, suggested that the MABC-2 is not a reliable assessment tool for children between the ages of 7 to 10. This is in line with previous research on age band-2

of MABC -2. Also, the TTS appeared to be more reliable than the respective subsections. The potential reasons behind these lower scores may be several, aside from the lack of reliability in itself. From the sampling perspective, the small and homogenous sample may have contributed to these results, particularly since correlation coefficients are sensitive to the lack of variance. Another potential reason for the restricted range may be the fact that some tasks were too easy (ceiling effect) or very difficult (floor effect). Also, the memory effect may have occurred as remembering what was done on test one inflated the scores on test two. In terms of the clinical implications, caution should be warranted when using the MABC-2 with children between the ages of 7 to 10. The standard scores were more reliable and produced more consistent results compared to the component and percentile scores. Also, the TTS should be used instead of, or in combination with the sub-component scores, to give a more reliable representation of a child's movement difficulties.

Acknowledgements

When I began this Masters degree, I often got asked the question "what are you going to do with that?" Although I still do not know the answer to that question, and I am still trying to figure out if there is a correct answer, I do know that I gained a plethora of knowledge, friendships, and life lessons while embarking on this crazy journey of graduate school.

From day one of grad school, we were constantly reminded that no matter how long you think something is going to take, it is actually going to take double that amount of time. Unfortunately, I found out the hard way that you should actually triple the time! Rule number one: TRIPLE THE AMOUNT OF TIME YOU THINK ANYTHING WILL TAKE!

There are many people that I have to thank for my success in grad school. Without each and every one of you I would not have survived, at least not with the little sanity that I do have left! I met Dr. Eryk Przysucha in my third year of undergrad where I was enrolled in many of the courses that he taught and after that there was no turning back! I consider myself very lucky to have had a supervisor that was wonderful through all stages of the grad school journey. Eryk saw me through my best and my worst moments and was always there to support me. Whether I was completing my proposal/defense presentations or I was in his office in tears because things just were not working out (thanks for always being a shoulder to cry on!), he always believed in me, even when I did not believe in myself. Thank you for pushing me to my limits in the world of academia and putting so much time and effort in to my work over the years. I appreciate everything you have done for me more than you will ever know. A wise supervisor once told me that if I learned anything from this experience, it should be that the sky is the limit, and this is something I will never forget!

I would also like to send a huge thank you to Dr. Carlos Zerpa and Mr. Vineet Johnson, two members of my committee. Without your critiques, knowledge and advice, my final document and presentations would not be at the level they are today. A special shout out to Carlos for all of his statistics expertise! Although sometimes your messages get "lost in translation", I could not have done this without you!

To the other graduate students that I was fortunate enough to share this experience with (you know who you are!), thank you! I did not expect to come out of this program with ten new, like-minded best friends, but you are all truly great people and I admire each and every one of you. Our weekly "venting" sessions were a necessity and it was so nice to have individuals around me who were all going through the exact same thing. Thank you for understanding why I had no social life throughout the years! I will always be here for all of you and am so proud of everyone for getting through this journey! Also to all of the other Kinesiology faculty and staff members, thank you for being those friendly, smiling faces that were always willing to listen or give advice when needed!

Lastly, to my family that supported me through everything. Thank you to my mom, aunt Joan, grandpa, and Nick. Without the constant encouragement and support from the four of you, I would not be where I am today! Mom, no matter what time of the day or night I was freaking out at, you were always just a phone call away and were always willing to talk to me and somehow knew exactly what I needed to hear. My grandpa always inspired me to be the best I could be and I wish that he could be here to see the finished product, but at least he was around for most of the journey and I got to hear the phrase "you're going to school all the way in Thunder Bay" many, many times! Thank you again to everyone that believed in me and was a part of my journey throughout these past three years!

Table of Contents

Key Definitions

Accuracy: The degree to which an observed score is in agreement with the true value and is free of measurement error.

Assessment: A process of collecting data through quantitative and qualitative methods, which allows an examiner to demonstrate an individuals' ability and identify the areas that need improvement. It can be formal or informal.

Bias: The manner of favouring one view or another, often not based on fair judgement, which results in invalid inferences.

Ceiling Effect: When a test cannot measure above a certain point because there is a distinct upper limit for responses. When most participants score near this upper limit the test is classified as being too easy.

Classical True Score Theory (CTT): A measurement theory that states that an observed score is comprised of a true score and measurement error. Although the true score will never be known, the smaller the error is then the closer the observed score reflects the true ability of the person. The reliability estimate is calculated as the ratio of the true variance to the total variance.

Consistency: Also known as stability, and it represents the degree to which the same results are achieved across trials or testing periods.

Correlation: A statistical procedure that measures the magnitude and direction of the linear relationship between two or more variables. The coefficient can range between -1 and +1, with a perfect relationship being indicated by a value of -1 or +1. The direction of the relationship is indicated by whether the coefficient is negative or positive.

Cronbach's Alpha: A statistical method used to estimate the internal consistency of a measure. Internal consistency is concerned with how well all the items in a test or instrument measure the same construct. The reliability coefficient estimates the correlation between the items and it can

range from 0 to +/-1. If the items within a test are correlated with each other, the value of alpha will be increased, hence the internal consistency will be high.

Floor Effect: The opposite to the ceiling effect. When a test has a distinct lower limit for responses and most participants' score at, or near, this lower limit. When most participants score near this lower limit the test is classified as being too hard.

Intraclass Correlation (ICC): A coefficient that is used to examine the relationship between two or more variables. It has only one distribution; hence it applies when the same people are tested across different and/or many conditions. Also known as an absolute measure of reliability, as it accounts for systematic bias.

Measurement: The process of assigning numbers to observations or individuals in a systematic manner as a way of representing or quantifying properties or characteristics of each individual.

Measurement Error: The degree of difference between a score achieved on a test and an individuals' actual ability. The errors can be random or systematic, and can result from many different sources and are unavoidable.

Normative Data: Information provided from a specific population that establishes baseline values for comparison purposes. Generally, this data is collected from a large, randomly selected, representative sample.

Psychometrics: A field of study that examines the characteristics of tests, such as reliability and validity. Psychometrics is concerned with the theory and technique performed to estimate the attributes of interest.

Reliability: A way of assessing the quality of the measurement procedure. Estimated as the degree to which the observed score is free of measurement error and accurately represents the true score.

Sample: A group of individuals that are selected to participate in the study. The sample is representative of the parameters within the study, which allows the results to be generalized from the sample to the theoretical population.

Standard Error of Measurement: An absolute measure of test reliability that allows for the estimation of random error. It can be used to create confidence intervals around the observed score of the person.

Test: A procedure that allows us to attempt to measure a certain characteristic of interest, whether it is intelligence, behaviour, or psychological status.

Test Retest Reliability: The degree to which a measure is consistent across different testing sessions. Estimated through the administration of the same test, to the same sample on two or more occasions, which are usually spaced a minimum of one week apart. The scores from these separate sessions are then correlated with each other.

Validity: The degree to which a test or instrument measures what it is intended to measure.

Variability: The degree to which the scores are distributed around the mean (intra-group vs. intra-individual). With a larger variability, the mean becomes less representative of the other scores. However, a lack of variability in the data also represents an issue, particularly in correlational analysis.

Introduction

Assessment

Assessment is the process of collecting and gathering information so that the assessor can infer an individual's achievement, skills, personality, and/or abilities (Losardo & Notari-Syverson, 2001). Within the field of motor behaviour, assessments can include any systematic method of obtaining information from tests and other sources that allow a professional to use this information to draw inferences about the movement produced (Losardo & Notari-Syverson, 2001). There are many different purposes of assessment. An assessment can be implemented to make the diagnosis of a movement difficulty. Also, it aids in the planning of programs for remediation and management, and it can be used to assess the effectiveness of interventions that may already be in place. Depending on the aim of the assessment process, there are different types of procedures that can be conducted to gather the information of interest.

Types of Assessments

There are different ways of classifying types of assessments, but generally a distinction is made between formal and informal approaches. Informal type of assessment often includes observations, interviews, and performance reviews that are less structured than formal assessments, and are likely less reliable and valid. Information that is gathered from informal assessments often aids in the examiners' initial decision as to whether or not to refer the individual for formal assessments. The most critical difference between formal and informal tests is that the former are presumed to exhibit a much higher degree of reliability and validity. These particular aspects of the assessment test are generally examined and re-examined across many psychometric studies. Formal assessments generally are structured and methodical, and aim to follow a set of guidelines, including specific administration techniques and scoring protocols.

These factors can affect the measurement error through non-uniform scoring guidelines, carelessness, and computational errors. Formal assessments generally involve standardized tests and can be further classified as criterion and norm-referenced tests. The former kind involves comparing achievements against objective reference points that have been clearly stated with criteria. On the other hand the norm-referenced tests rely on quantitative data gathered from a larger population. Normative data allows a researcher to compare results from a sample to results that are characterized as "normal" in a defined population and draw inferences about the execution of the task(s) or assessment in relation to a larger population. An example of a formal, normative test is the Movement Assessment Battery for Children – First and Second Editions. Both versions of this test represent an assessment tool that is standardized and often used by professionals to identify mild to moderate movement difficulties in children (e.g., Developmental Coordination Disorder (DCD)) (Barnett & Henderson, 1998). The first version of this test has been considered by many clinicians and researchers as a gold standard in the area, however the psychometric qualities of the second version are still equivocal.

Theories of Measurement

Measurement error is present in every type of assessment, whether formal or informal. The Classical True Score Theory (CTT), which is one of many theories of measurement (e.g. Generalizability Theory, Item-Response Theory), is able to estimate how much error is within the measurement (Novick, 1966). The goal of CTT is to establish the reliability of an assessment and this is measured through the use of many different estimates (Suen & Lei, 2007). There are assumptions that must be met in order to proceed with the analysis to ensure that the results are correct. The first assumption that must be met for the CTT to be reasonable is that the participant's observed score must equal his/her true score plus error (Lord & Novick, 1968). The

true score component is the value that would be obtained if there was no measurement error; however, this is not possible. One approach to estimate the true score is by taking the mean of an infinite number of trials performed by the same individual, under the same circumstances, and on the same test. Unfortunately, this method is impractical. Thus, since the true score cannot be inferred directly, it must be estimated from the observed score. According to CTT, the only way that the observed score can be an accurate estimate of the true score is when the amount of measurement error (random and systematic) is small. Another assumption that must be met is that the expected value of any observed score is the persons' true score. Therefore, the mean of all the errors would be equal to zero (McDowell, 2006). As well, the covariance of error and true score components of a person's observed score must be zero for a population, which demonstrates that the random errors are uncorrelated with the true scores. This assumption implies that there is no systematic relationship between the true score, and whether the persons' errors are positive or negative (Shultz, 2005).

Measurement error may be attributed to a variety of factors that systematically or randomly affect the scores. Systematic measurement error regularly affects an individuals' score and is due to a specific characteristic of the person, the test, or the environment (Crocker & Algina, 1986). These characteristics then cause the observed score to be an inaccurate representation of the persons' true abilities. Unlike random error, the systematic error can be removed. Random error, often called sampling error, affects an individual's score due to trial-to-trial variability, which could include, but is not limited to, environmental factors (e.g. location, time of day, and/or social factors), individual (e.g. guessing, an individuals' mood, inattention, tiredness, and/or misunderstanding of directions) and demographic factors (e.g. age and/or sex) (McDowell, 2006).

Reliability and Its' Estimates

In order to be confident in results obtained by an assessment or test, it is important that that assessment or test has strong psychometric properties, and therefore the inferences emerging from the results would be valid. Psychometrics focuses on the examination of the reliability and validity of an assessment. Reliability is of particular importance because without reliability, the validity is impossible to achieve. Although many definitions can be used, here reliability is operationalized as the degree to which a score is free of measurement error (Weir, 2005). High reliability indicates that the emerging score on a test is in the absence, or with minimal amount of measurement error (Atkinson & Nevill, 1998). If there were a lot of measurement error, that means that if the individual was to be re-tested again and again, his/her observed score pertaining to a relatively stable trait would potentially fluctuate from one testing session to the next. Among many different approaches, reliability is often estimated through test re-test, internal consistency, and SEM.

Test Re-test. One way of inferring the amount of the emerging measurement error is through test retest approach, which examines the consistency/stability of repeated performances that are separated in time and measured by the same examiner under the same conditions (Furr & Bacharach, 2008). Ideally, the test would be re-administered two to four weeks apart (Furr & Bacharach, 2008). If the time elapsed between testing sessions is too short, the practice effect might have an impact on the results. The participants may remember the tasks that they must complete and could even have been practicing them for the time between testing sessions. Therefore, the scores on time two would be inflated and would not constitute an accurate representation of the participant's abilities. On the other hand, if the testing sessions are taking place far apart, the maturation effect may have an impact. Children are constantly learning and

growing, and because of this fact there could be a large mental and/or physical change in the child, potentially affecting the results. Due to these factors, it is important to administer a second testing session within two weeks time. Generally, test re-test reliability is higher when the time span between test administrations is shorter rather than longer (Miller, 2008).

As with all types of reliability, there are assumptions that must be met with test-retest approach. Conceptually, the attribute of interest should be stable and consistent across time one and time two. Hence, it is assumed that if other factors are controlled for (e.g., maturity; practice; memory effect; consistency in protocol) the attribute of interest (e.g., motor proficiency) should remain stable. Also, the error variance must be equal between the first and the second testing session, which can be done by doing a paired t-test to infer if there were changes in the mean (Furr & Bacharach, 2008). If these assumptions are met, the correlation between the scores from trial one and trial two are an accurate estimate of the reliability.

Intra-class correlation (ICC) is widely used to examine the relationship between a variable across two or more testing sessions (McGraw & Wong, 1996). The ICC reflects the degree of consistency, or the absolute agreement, between two or more sets of data collected on the same sample/population (Bruton, Conway, & Holgate, 2000; Kim, 2013). There are at least 6 different versions of the ICC, all of which can give quite different results even when applied to the same data (Atkinson & Nevill, 1998). To choose an appropriate version of ICC, one must understand how the data has been collected and what type of question is being examined. For this study, a two-way mixed model with absolute agreement was the type of ICC used for analysis. To ensure that testing conditions remained the same, or as similar as possible, the same examiner assessed the participants on each occasion, and the examiner was not chosen randomly from a larger properly. This avoided the possibility of inter-tester bias.

Most commonly, ICC is based on the calculation of the F-value from the repeated measures ANOVA and examines the consistency for pairs, or sets of measurements. The interpretation of scores is dependent on the context of the data. A correlation coefficient of 0.8 demonstrates a good reliability, however it is important to remember that this value is context specific. For example, if the correlation coefficient was 0.7 with 500 people, this would not represent a high reliability, however, if the coefficient was 0.7 and there were only 50 people, this correlation may be considered as high. A coefficient of around 0.9 may be needed to achieve a high reliability for a larger population. Generally, an ICC value that ranges between 0.7 and 0.8 is deemed to have moderate reliability, and a value greater than 0.9 is indicative of high reliability (Atkinson & Nevill, 1998). Thus, if the absolute scores remain relatively the same across both tests for each person, and given that the data set is not homogeneous, then a high ICC would emerge.

Another coefficient that is used in test re-test studies is the Pearson Product-Moment Correlation, which is a bivariate measure of association between two date sets. Similarly to ICC, the Pearson correlation coefficient can range from -1 to +1, with a value greater than 0.8 representing a high reliability (Atkinson & Nevill, 1988). The main advantage of using ICC over Pearson Product correlation is that, as previously stated, ICC is a univariate rather than a bivariate measure. Univariate analysis is generally more reliable since only one variable is being used and this method can be used to compare a test with multiple re-tests (Atkinson &Nevill, 1998). Another advantage of ICC is that it is sensitive to the presence of systematic error in the data. The ICC decreases in response to both lower correlations between raters, in the case of inter-rater reliability estimate, and larger mean differences in the case of the test-retest approach. Pearson may produce a high reliability coefficient, indicating consistency. Yet, in absolute terms

the mean performance across the two testing sessions may be significantly different, as inferred from the analysis of variance (e.g. dependent samples t-test). Thus, ICC accounts for the absolute consistency across the scores (Osborne, 2008). This is why ICC is known as an absolute measure of reliability, compared to Pearson's correlation, which is a relative measure.

There are some statistical threats that may affect ICC and/or Pearson's, and therefore affecting the reliability coefficient. The size and characteristics of the sample represent such factors as small and homogeneous data sets may coincide with a lack of variability, therefore artificially deflating the reliability coefficient (Baldwin et al., 2011). With a larger sample size, there is a better chance that the data is normally distributed, as it is more heterogeneous. Hence, when smaller samples are involved it is important to examine how/if at all the normality of the data set is negatively affected (Weir, 2005). The magnitude of a correlation may also be affected when the data set is too heterogeneous, as it is the case when the outliers are present. Outliers can result from faulty sampling design, data collection errors, or the fact that the subject does not understand the task at hand (Goodwin & Leech, 2006). The effect of an outlier is greater when the sample size is small, and it can substantially decrease the magnitude of the correlation (Goodwin & Leech, 2006). Generally, the presence of outliers can be detected by visually inspecting the scatter-plots of the respective data sets.

Internal Consistency. The second type of reliability used here is internal consistency. This type of reliability assesses how well the items of a test or instrument measure a specific construct (Furr & Bacharach, 2008). If the individual items within the sub-sections of a test are highly correlated with one another, the estimate of reliability will be higher (Kirk & Miller, 1986). Having a high correlation amongst the items would indicate that the items represent the same construct, and yield similar results. For example, the manual dexterity sub-component of

the MABC-2 consists of three individual tasks. If the three items within the sub-component have a high internal consistency, this is an indication that all three tasks measure the domain of manual dexterity. This form of reliability only requires one test administration and the estimations are achieved through split-half reliability and coefficient alpha (Furr & Bacharach, 2008). For the purpose of this study, Cronbach's alpha will be implemented.

Coefficient alpha, often called Cronbach's alpha, is the most common internal consistency measure and has more practical advantages than any other methods (Furr & Bacharach, 2008). When computing a raw coefficient alpha, one must first obtain a set of item-level statistics. The variance of scores on the complete test are then calculated and followed by the calculation of the co-variances between each pair of items (Furr & Bacharach, 2008). If two items are reliable measures of the same construct, they should have a positive co-variance value. If the value is not positive, this makes the examiner aware that the items either do not measure the same construct, or at least one item is affected by measurement error. Co-variances then need to be summed, which produces a value that shows the degree to which responses to all of the items are consistent with each other (Furr & Bacharach, 2008). When the sum of the co-variances is large, the items are found to be more consistent with each other. The estimate of reliability can then be calculated by entering the variance of the scores on the complete test and the sum of the co-variances into an equation, which takes into account the number of items that are in the test.

In comparison to the raw coefficient alpha, another method used to estimate the internal consistency is the standardized coefficient alpha. The main difference between the two methods is that the latter method uses test scores where all items have been standardized before calculating the estimate of reliability (Furr & Bacharach, 2008). For the purpose of this study, the

raw coefficient alpha will be used for analysis because there is no need to standardize the data as it is all based on the same scale. Had there been a variety of scales used (e.g. dichotomous vs. continuous), it would have been more appropriate to use the standardized coefficient alpha.

There are potential threats to the internal consistency reliability estimate. It is important that all components of a test are consistent, measuring similar constructs if they are in the same sub-component, to ensure the test as a whole has a greater reliability (Furr & Bacharach, 2008). In the present study, the internal consistency was assessed using Cronbach's alpha, which examined the consistency of the eight individual items, in their respective sub-sections, within the assessment tool. The length of a test and number of items within the test are two factors that also affect the value of coefficient alpha. A longer test will provide higher values of Cronbach's alpha and is more reliable than a shorter test because of the higher degree of variance and its' relation to reliability (Tavakol & Dennick, 2011). As the length of the test increases, the observed score variance increases at a faster rate than the error score variance (Furr & Bacharach, 2008). Thus, since Cronbach's alpha is calculated as a function of the number of test items and the average inter correlation among the item. If the number of items is increased, Cronbach's alpha is also increased (Tavakol & Dennick, 2011).

Standard Error of Measurement. The standard error of measurement (SEM) estimates the degree to which an observed score would very if the person was to be re-tested over and over again. The SEM is a measure of absolute reliability. This estimate quantifies the reliability of scores within an individual on different testing sessions (Overend, Anderson, Sawant, Perryman, & Locking-Cusolito, 2010). Thus, the benefit of calculating the SEM, as opposed to other estimates of reliability, is that it allows a researcher to make statements about the precision of the test score of an individual rather than the test (items) itself (Harvill, 1991). The SEM is based on

the estimated reliability coefficient, which in this case is ICC from the test re-test reliability (Suen & Lei, 2007). The SEM is calculated based on the following equation: SEM = SD √ (1 – r), where SD represents the standard deviation of the errors of measurement and r is the reliability coefficient. Thus, as evident from the equation, the magnitude of SEM depends on the variability of the sample and/or the magnitude of the reliability coefficient. Generally, when the variability of the group is high and the reliability of the measure low, the resulting value will be high, and vice-versa. A high reliability coefficient and low SD would result in a low SEM. Once calculated, the SEM can be used to form confidence intervals around the observed score

Confidence intervals have lower and upper limits, which are determined based on a range of scores that have a high probability of including an individual's obtained score, if he/she was to be re-tested. The most typical confidence intervals established are 68%, 95%, and 99%, and are chosen based on the level of confidence that one would want to have. If a 68% confidence interval were chosen, which represents one SD, this would produce the smallest range of scores, as compared to the 95% (two SD) or 99% (three SD) confidence interval. For example, if the observed score was 15, and the SEM was ± 2, this means that one can be 68% confident that the score would range from 13-17, if the individual was to be tested again. Increasing the interval to 95%, allows for more confidence that the obtained score would fall within the provided range. However, since the range of scores is now much wider and less precise, thus the obtained score could still be far away from the true (hypothetical) score.

Tests of Motor Impairment

Reliability issues are of particular interest to the professionals in the adapted physical activity field. One important role of an adapted professional is to assess individuals and identify if they have physical, social, or psychological problems (Ellinoudis, Evaggelinou, Kourtessis,

Konstantinidou, Venetsanou, & Kambas, 2011). Adapted professionals must rely on formal assessments to make valid judgments about the attributes of interests. If those tests lack reliability, the results may lead to false positive or negative inferences, which may have detrimental consequences for the client. As it is impossible to have valid inferences without reliable tests, the latter issue is of primary interest here. The two most frequently used formal, norm-references tests are Bruininks-Oseretsky Test of Motor Proficiency (BOTMP 1 and 2) and MABC (1 and 2), with the latter being the main focus of this study.

Bruinkins-Oseretsky Test of Motor Proficiency. The BOTMP measures both fine and gross motor skills (Bruininks, 1978). It is used for the purpose of screening, evaluation, research, program planning, and assistance with placement decisions (Wiart & Darrah, 2001). Generally, children enjoy completing this assessment as there are a great variety of test items and they often are appealing for the designated age group. There are eight subsections within the BOTMP including fine motor precision, fine motor integration, manual dexterity, bilateral coordination, balance, running speed and agility, upper-limb coordination, and strength. Within each of these subsections, the number of activities ranges from five to nine, with a total of 46 items making up the complete battery (Wiart & Darrah, 2001). Completion of the tool provides composite scores in four motor areas and one overall motor-proficiency measure. The sub-sections of the assessment are fine manual control, manual coordination, body coordination, and strength and agility. This tool is designed for children between the ages of 4 years 6 months to 14 years 5 months (Venetsanou, Kambas, Aggeloussis, Serbezis, & Taxildaris, 2007). In terms of its reliability, Moore and colleagues (1986) examined the test re-test reliability of the BOTMP. Thirty-two, 5 year old children, were tested on two occasions, one week apart. The results

revealed an ICC coefficient of 0.76 for test re-test reliability of the composite score. The reliability ranged from 0 to 0.76 for the sub-sections of the assessment tool.

Recently a revised version of the assessment tool was released, BOTMP-2 (Bruininks & Bruininks, 2005). Wuang and Su (2009) examined some of the psychometric properties of the BOTMP – second edition. One hundred atypically functioning children between the ages of 4 and 12 were tested on 3 occasions, two baseline measurements were recorded two weeks before the intervention, and a four-month follow up measurement took place after the completion of a paediatric rehabilitation program. Test re-test reliability, assessed using ICC, produced a coefficient of 0.99 for the total motor composite score, indicating excellent reliability (Wuang & Su, 2009). For the four sub-component composite scores, the ICC for test re-test ranged between 0.88 and 0.99 (Wuang & Su, 2009). The researchers also reported a Cronbach's alpha of 0.92 for the internal consistency of the total motor score, which also shows excellent reliability. The alpha values for the sub-components ranged between 0.78 and 0.97 (Wuang & Su, 2009).

The BOT-1 and BOT-2 are considered to be gold standards for formal assessments in the field of adapted physical activity, as is the MABC (Slee, Campbell, & Spears, 2012; Henderson, Sugden, & Barnett, 2007). There are some similarities between these two assessment tools and MABC/MABC-2. The BOT-1/BOT-2 and MABC/MABC-2 both are designed to provide a comprehensive assessment of motor development and proficiency. As a result, one of the sub-components (manual dexterity, upper-limb coordination, and balance) are similar in both tests. As well, both assessment tools provide composite, standard, and percentile scores. The BOT-1/BOT-2 provides a Total Motor Composite score, which summarizes the overall results of the assessment. This score is similar to the Total Test Score (TTS) implemented in MABC.

Movement Assessment Battery for Children (MABC). Although the choice of a particular test depends on many factors (e.g., purpose; population of interest; theoretical framework), MABC has been considered as a gold standard in the area of adapted physical activity as related to the assessment of children with non-congenital, developmental coordination problems. The MABC is also a standardized, norm-referenced assessment test which is often used by professionals to identify mild to moderate movement difficulties in children (e.g., DCD) (Barnett &Henderson, 1998). As it was the case with BOTMP, the performance test involves the completion of fine and gross motor tasks, categorized into manual dexterity, ball skills, and balance (Wiart & Darrah, 2001). The assessment includes eight items in total, and together these items make up the Total Impairment Score (TIS). For example, in age band 2, the manual dexterity section contains three different tasks including shifting pegs by rows, threading nuts on a bolt, and the flower trail. There are two tasks for the aiming and catching section. These tasks are a two handed catch and throwing a beanbag in to a box. The last sub-component, balance, includes one-board balance, hopping in squares, and ball balance tasks.

MABC: Old vs. New. Recently, a revised version of MABC has been designed (Movement Assessment Battery for Children – Second Edition) (Henderson, Sugden, & Barnett, 2007). Revisions included making the "kit easier to carry, the performance test items are more engaging for children, and the scoring system for both the performance test and checklist are more user-friendly" (Brown & Lalor, 2009, p. 92). The reason for the revision was to enhance the tool and provide an updated version that was more easily administered, while maintaining the reliability and validity of the items (Henderson et al., 2007). The new version of MABC encompasses a broader age range than the previous test. Children and adolescents aged 3 through 17 can now be assessed, in contrast to previous age bracket (4 through 12) (Brown & Lalor,

2009). Also, the number of age bands has been reduced from four to three (3 to 7; 7 to 10 and 11 to 17). In addition, a number of items were revised, and new equipment was introduced (see Appendix A). Focusing on age band 2, in the manual dexterity subsection, the placing pegs task now has a new starting position and layout. The lacing board is longer for the threading lace activity, and the shape of the drawing trail has changed (Brown & Lalor, 2009). For the aiming and catching section, in the beanbag task a box was replaced with a target, whereas in the balance subsection floor mats were added for the one leg hopping activity. Despite these numerous changes, the authors maintained that the research pertaining to the reliability and validity of the old tool applies to the new version. However, others questioned that assumption and called for more investigations examining the respective issues with the new version of MABC (Brown & Lalor, 2009).

Psychometric Properties of MABC-2. Venetsanou and colleagues (2011) completed a literature search for articles regarding the original version of MABC test to determine if the assessment should be considered as a "gold standard". The authors found only five studies, which explicitly examined the reliability of the original version, but the inferences still remain equivocal. Croce and colleagues (2001) reported strong reliability for the composite and sub-scores, but the analysis combined three different age groups (age band 1, 2 and 3). As a result, the emerging inferences warrant caution and are difficult to interpret if one is interested in one particular set of items. In fact, when Chow and Henderson (2003) conducted a similar study examining the test re-test reliability of age band 1 alone, the reliability coefficients were substantially lower reflecting moderate to weak reliability. In terms of the new version even fewer studies have been conducted. Most of the existing research involved age band 1 (3 to 7 years old) (Ellinoudis et al., 2011; Hua, Wu, Gu, & Meng, 2012; Hua, Gu, Meng, & Wu, 2013;

Smits-Engelsman, Neimeijer, & van Waelvelde, 2011), with only one study examining age band 2 (Holm, Tveter, Aulie, & Stuge, 2013) and age band 3 (Chow, Chow, Chan, & Lau, 2002).

Age Band 1. Smits-Engelsman and colleagues (2011) explored the reliability of MABC-2 in 3-year olds. Fifty typically functioning children, between 3 and 4 years of age, were assessed individually with one to two weeks apart. . Using the component scores, test retest reliability was measured with ICC, and internal consistency was measured with Cronbach's alpha. The ICC values for manual dexterity, aiming and catching, and balance sub-sections were 0.85, 0.74, and 0.75, respectively, which indicated that the sub-sections of the test were reliable. The ICC for the total test score was 0.83, also indicating good reliability. The internal consistency was calculated based on the 10 item standard scores, as both the left and right scores during the placing pegs task and one-board balance tasks were used. Cronbach's alpha values were 0.81 on the first testing occasion and 0.87 on the second testing session. Thus, the values showed that the internal consistency was acceptable to good. The SEM was also calculated for the TTS and sub-components to determine the precision of the total test sore, with the results ranging from 0.73 to 1.47.

In a larger study, Ellinoudis and colleagues (2011) examined 183 typically functioning children (98 males and 85 females) between 3 and 5 years of age. The reliability was examined using test retest and internal consistency estimates. The children were assessed individually twice, one week apart. The ICC coefficient for test re-test showed that the reliability of the individual items, other than the trail drawing activity, was moderate to excellent, varying between 0.73-0.96. The drawing trail activity, which was reported as the problematic task, achieved an ICC of 0.66, which is considered to be approaching a moderate degree of reliability. The sub-sections of manual dexterity, aiming and catching, and balance had ICC values of 0.82,

0.61 and 0.90, respectively. The TTS had an ICC value of 0.85, indicating good reliability. The internal consistency was by examined with Cronbach's alpha for the items within each of the three domains. For the three subsections of manual dexterity, aiming and catching, and balance, the coefficient values were 0.51, 0.70, and 0.66, respectively. These values indicated moderate to poor internal consistency for the respective subsections. The authors suggested that the moderate to low internal consistencies could be due to the relatively small number of items within each domain of the assessment (i.e. 2 tasks for aiming and catching). The raw scores were used for the analysis of reliability at the individual item level of the test, whereas the standard scores were used for the 3 sub-sections and TTS scores.

Most recently, Hua and colleagues (2013) examined the reliability of age band 1 with 184 children, between 6 and 8 years of age. The ICC was used to examine the inter-rater and test re-test reliability, and Cronbach's alpha was once again selected to analyze the internal consistency of the MABC-2. Inter-rater and test re-test reliability incorporated the raw scores for analysis, while internal consistency involved the standard scores for the eight items and the total test score. Internal consistency was not calculated for the sub-components, but instead was calculated for each of the eight items. The alpha values ranged from 0.23 to 0.60, indicating a weak to questionable internal consistency (Hua et al., 2013). Two of the eight items, drawing trail and walking heels raised, had considerably lower values than the other tasks (0.23 and 0.27, respectively). There was no explanation as to why these two scores were much lower. However, when they were both deleted the Cronbach's alpha coefficient increased substantially. This is an indication that these two tasks were problematic. Cronbach's alpha for all eight items was 0.50, indicating a poor, but acceptable, internal consistency. The ICC for intra-rater and test retest

reliability (based on the individual scores) was found to be excellent, with values close to or above 0.90 for the respective items.

Age Band 3. Chow and colleagues (2002) examined an experimental version of the MABC-2 with focus on age band 3. The examiners used 31 adolescents ranging between 11 and 16 years of age to examine inter-rater and test retest reliability. The ICC coefficients varied from 0.92 to 1.00 for inter-rater reliability, and 0.62 to 0.92 for test retest reliability using the total test score. In this study, the researchers used the 15th percentile as a means of categorizing children into impaired or not impaired categories. This means that if a child's score was below the 15th percentile on both testing occasions, then the scores were considered to be perfectly reliable, perhaps inflating the ICC.

Age Band 2. In terms of age-band 2, which is of primary interest here, only one study has been conducted so far. Holm and colleagues (2013) examined the reliability of age band 2 using intra and inter-rater reliability coefficients. Forty-five typically functioning children, 23 girls and 22 boys, with a mean age of 8.7 years were recruited. Inter-and intra-rater reliability were evaluated using ICC. When the children attended the first testing session, they were tested twice by two physiotherapists, who scored them independently (inter-rater). On the second testing session, one to two weeks later, the children were re-assessed by one of the examiners (intra-rater). The analysis of the component scores showed that there was a lack of reliability for the sub-sections as well as the TTS. Intra-rater reliability had ICC values of 0.62, 0.49, and 0.49 for manual dexterity, aiming and catching, and balance respectively. The ICC for the TTS was also low 0.68, and the SEM for the different scores were 3.2 (manual dexterity), 2.4 (aiming and catching) 2.7 (balance), and 4.9 (TIS). The analysis of inter-rater reliability, also calculated based on the component scores, had ICC values of 0.63, 0.77 and 0.29 for manual dexterity,

aiming and catching, and balance. As well, the ICC for the TTS was 0.62, indicating a questionable degree of reliability. The SEM reported for manual dexterity (3.2), aiming (2.0) and catching, balance (4.5) and the TTS (6.8) were also large. Among the individual items, the threading lace and one board balance tasks had the highest SEM for both intra-and inter rater reliability (i.e. 4.7 and 5.3 for intra-rater, and 4.1 and 7.3 for inter-rater). Holm and colleagues (2013) suggested that these two tasks might be too challenging for the participants, therefore creating a ceiling effect. Overall, this study showed that across the different scores the reliability of the test was moderate to questionable. Also, it revealed that certain tasks within the MABC-2 (e.g. tasks within the balance sub-component) maybe problematic, as they are too difficult or not challenging enough. The important limitation of this research was that the researchers only used inter and intra-rater reliability, thus the internal consistency of the MABC-2 was not examined. As well, the researchers only analyzed the component scores and did not address the reliability of the other scores (standard & percentile), which are often used in research and clinical settings.

Purpose

The purpose of this study was to examine different facets of reliability (test-retest, internal consistency, and standard error of measurement) of the MABC-2, age band 2 (7 to 10 years old) across standard, component, and percentile scores for the three subsection and TIS scores.

Hypotheses

1. Test re-test reliability (ICC) for the TTS would be moderate to high (> 0.70) for the standard, component and percentile scores.

2. All three sub-sections (manual dexterity, aiming and catching, and balance) would demonstrate a moderate to high test re-test reliability across trials for the standard, component, and percentile scores.
3. Internal consistency (Cronbach's alpha) would be moderate (> 0.70) for manual dexterity and aiming and catching sub-sections.
4. The internal consistency (Cronbach's alpha) would be questionable (< 0.60) for the balance sub-section, due to the inconsistencies of the one-board balance task.
5. The SEM for the TTS and three sub-sections (manual dexterity, aiming and catching, and balance) would be the lowest for the standard scores, in comparison to the component and percentile scores.

Method

Participants

Forty participants between 7 and 10 years of age (ME = 9 years, 0 months and 5 days) were recruited for this study. Both males and females participated (18 males and 22 females). In order to be included in this study, children were required to be typically functioning in terms of their motor and cognitive status, as reported by the parents/guardians. Atypically functioning children with an official diagnosis for any developmental disabilities in the cognitive or motor domains were excluded from the study.

Recruitment

Purposive sampling was implemented. The recruitment process was initiated by submitting an application to the director of education for the Thunder Bay Catholic District School Board and the Lakehead Public School Board in Thunder Bay, Ontario (Appendix B and C). Both boards agreed to cooperate with the recruitment, and the packages were delivered to the teachers of the appropriate grades. The students were asked to return the forms to their teacher the following week and the packages were then picked up. The participants were also recruited from local soccer teams and through the word of mouth. The recruitment packages were handed out to all parents whose children were a part of Lakehead Express U-10 Soccer Club. They were asked to return the forms to the coaches the following week, if interested. The recruitment package contained a recruitment letter, consent form, and an ExPARA (Exercise and Physical Activity Readiness Assessment) (Appendix D, E, and F, respectively). Consent forms required the parents' signature as the children were all under the age of 18. When the forms were returned, parents were contacted via phone or email to set up the testing sessions. Prior to the

data collection, participants were given a brief description of the study. The children and parents were all informed that participation was voluntary and that all data would remain confidential.

Procedures

Participants were asked to commit to two sessions, one to two weeks apart. Children were assessed individually and each session took approximately 45 minutes to one hour. At the second testing session, the child was re-assessed at the same location, under the same conditions, and by the same examiner.

There are 8 different tasks in the MABC-2 test for age band 2, which are divided into three sub-sections; manual dexterity, ball skills and balance. The tasks within each section are safe, relatively simple and resemble activities that a child performs on a daily basis, either in school or on the playground. There are 3 manual dexterity tasks (placing pegs, threading lace, and the drawing trail-2), 2 ball skills tasks (catching with two hands and throwing beanbags onto a mat), and 3 balance tasks (one board balance, walking heel-to-toe, and hopping on mats).

For each of the manual dexterity tasks, a demonstration was given at the beginning and then the child was allowed to complete one practice attempt to ensure their understanding of the task. To complete the placing pegs task, the child was asked to pick up the pegs, one at a time, and insert them into the board as quickly as possible, as this was a timed event. The timing stopped when all 12 pegs had been placed into the peg-board. This task was done with both the preferred and non-preferred hand. The threading lace activity was another timed, manual dexterity task. The child was asked to pick up the lacing board and insert the lace through the first hole and then continue threading the lace back and forth in a straight line through the remaining holes, as quickly as possible. The last of the three manual dexterity tasks was the drawing trail 2. This task required the child to draw a single continuous line following the trail

without crossing the boundaries. Only the preferred hand was tested for this task and the child was not timed, rather the number of times the line was crossed was recorded as an error.

Catching with two hands was the first of two tasks in the aiming and catching subsection. For this task, the child was required to throw a tennis ball at the wall from a distance of 2 metres, and then catch the ball with two hands when it bounces off the wall. If the child was 7 and 8 years of age, one bounce was permitted, however no bounces were allowed for the 9 and 10 year olds. Five practice attempts were given for this task and 10 formal trials. The number of balls that the child caught correctly out of 10 attempts was recorded. Secondly, the child was required to throw beanbags onto a mat. From a distance of 1.8 metres, the child stood on one mat and threw the beanbag to the other mat, so that the beanbag landed on the red circle in the middle of the mat. Ten formal trials were given and the number of beanbags landing on the centre red circle was recorded. If the beanbag was thrown and landed on the circle and proceeds to bounce off, this did not count as a successful throw.

The last subsection of MABC-2 is balance. The first task required the child to balance on one foot on the balance board, for a maximum of 30 seconds. Both legs were tested for this activity. The next task focused on dynamic balance and required the child to walk heel-to-toe forward on a 4.5 metre straight line that had been marked on the floor with tape. With the toe behind the starting line, the child had to place the heel of one foot against the toe of the other and continue walking on the line. The number of steps made to reach the end of the line was recorded. The last balance task required the child to hop on mats. Six mats were placed adjacent to one another in a row and the child started by standing on one foot on the first mat. Remaining on that same foot throughout the duration of the task, the child was asked to make 5 continuous hops from mat to mat, stopping on the last mat and maintaining his/her balance in a controlled

position. Both legs were tested, and the number of hops completed was recorded. For each child, a Total Test Score (TTS) and three sub-section scores were derived as raw scores were converted to standard, component, and percentile scores.

Measures, Design and Analysis

The reliability of the MABC-2 was analyzed using test re-test approach, internal consistency and standard error of measurement (SEM). The ICC was incorporated in the test-retest whereas the Cronbach's alpha was used to infer internal consistency. A repeated measure design was implemented, with time (pre vs. post-test) as the independent variable. In terms of the dependent measures, three types of reliability coefficients were derived from the data, for the three different types of scores. A two-way mixed model with absolute agreement was implemented for the calculation of ICC. An analysis of variance (dependent samples t-tests) was also implemented to determine if there were statistically significant differences between the group means across the two sessions. Also, scatterplots were generated to examine the homoscedasticity of the data, presence of the outliers, and the shape of the emerging data distribution.

With each type of analysis, there are necessary assumptions that the data must meet before further examination can happen. For ICC, the data must be normally distributed which would indicate that the mean, median and mode are similar and coincide with the peak of the bell shaped curve. In order to determine which set of data (standard, component, and percentile) was normally distributed, a Shapiro-Wilks test was implemented. If the significance value for the test statistic was greater than 0.05, the data was normally distributed. However, if the data was less than 0.05, this means that the distribution significantly deviated from the acceptable degree of normality.

The skewness and kurtosis of each data set was also examined. The analysis of the skewness of the distribution provides three different values, a test statistic, the standard error, and the normal distribution. If the test statistic is between – 0.5 and + 0.5, this indicates that the data is approximately normally distributed. If the statistic was greater/less than – 1/+ 1 then the distribution was highly skewed. If the standard error value were more than double the absolute values of the test statistic, this would be an indication that the data was not normally distribution. Lastly, a value representing the normal distribution is provided. The calculated normal distribution for standard, component and percentile scores should be around +/-1.00, if normally distributed.

Similar to the skewness analysis, the kurtosis results provided a test statistic, standard error, and normal distribution. The standard error and normal distribution are analyzed the same way as for skewness, which was discussed above. The test statistic for kurtosis determines whether the data is platykurtic, mesokurtic, or leptokurtic. If the value were less than zero, this would indicate that the distribution is platykurtic, which means that the central peak is lower and broader, compared to the normal distribution. A test statistic above zero reveals a leptokurtic distribution, meaning that there were more data points in the tails than around the mean.

Another assumption that must be met is that the data must be homoscedastic. Hence, the data must have equal scatter around the hypothetical line of best fit, indicating a similar variance across the data. Here, the homoscedasticity was assessed using scatterplots to determine if the data points were spread equally around hypothetical line of best fit, or if they were clustered to one area (e.g. TTS in Figure 1). There also must be a minimal number of outliers in the data set, which is implied when the data is homoscedastic.

Internal consistency was measured for the items in the sub-sections of manual dexterity, aiming and catching, and balance. The scores from time one were used to analyze the internal consistency of the three sub-components. As the TTS is produced from the addition of the scores from the sub-component categories, the internal consistency of the TTS was not examined in this study. Instead, each individual item within the three sub-components was examined to see if the tasks in each domain measured what it was intended to measure. Also, the Cronbach's alpha values were examined when the items were deleted. This allows inferring if the internal consistency of the sub-component improves (i.e. increases) when a certain item is removed. When an item is removed and the internal consistency increases this indicates that that specific task may be problematic within that domain. If the Cronbach's alpha stays the same or goes down, while the item is deleted, this indicates that the item enhances the internal consistency of the sub-component. The aiming and catching sub-component is composed of only two tasks and therefore the data cannot be further analyzed with items deleted approach. The SEM was calculated using the following formula: $SEM = SD\sqrt{(1 - r)}$ (Harvill, 1991), where SD is the standard deviation of the sample, and r is the reliability coefficient, in this case ICC.

Results

Tests of Assumptions

Previous research of MABC-2 failed to test the normality of the data. This is a reason for concern because it is plausible that some scores may meet the necessary assumptions, while others may not. None of the previous studies examined the percentile scores and instead focused on either the standard or component scores.

Normality, Skewness, and Kurtosis.

Standard Scores. The first assumption that was examined was the normality of distributions, via Shapiro-Wilks test. The results showed that three out of the eight values were significant at $p < 0.05$ or lower, indicating that these domains were not normally distributed. Balance control emerge as the domain which had questionable normality at both time one and time two, and the aiming and catching sub-component for time one. Skewness and kurtosis were used as another means of testing for normality. All domains had skewness statistics between -0.5 and 0.5, therefore indicating that the data was approximately symmetric. The standard error for both the skewness and kurtosis outputs was also provided (Table 1). If this value was more than double the absolute value of the test statistic, this is indication that the data is not normally distributed (Martin & Larson, 2006). As evident, the standard error for the skewness of the standard scores was 0.37 for all of the domains, and all the skewness statistics were within two standard errors, suggesting that the data was normally distributed (see Table 1).

Seven out of the eight kurtosis statistics were below zero and therefore representing a platykurtic distribution, where the central peak is lower and broader as compared to a normal distribution. As all of the values were below, but close to 0, this indicated that the distribution was marginally platykurtic. Similarly to the skewness results, all of the kurtosis statistics were within two units of the standard error (0.73), and therefore it can be inferred that the data was normally distributed. The kurtosis statistic for the total test score from time one was the only positive value out of the eight domains. Although the kurtosis statistics were similar across time one and time two, there was one pattern that emerged across the testing sessions. For all domains, the kurtosis statistics were lower and therefore more platykurtic, on time two than time one.

Table 1.

Results for Tests of Normality, Skewness and Kurtosis for Standard Scores Across First (T1) and Second (T2) Testing Sessions.

Standard Scores	Normality (Shapiro-Wilks)		Skewness			Kurtosis		
	Statistic	Sig.	Statistic	SE	Normal Distribution	Statistic	SE	Normal Distribution
Total Test Score (T1)	0.97	0.34	-0.17	0.37	±0.45	0.45	0.73	±0.61
Total Test Score (T2)	0.96	0.22	0.08	0.37	±0.20	-0.06	0.73	±0.08
Manual Dexterity (T1)	0.97	0.35	-0.14	0.37	±0.37	-0.42	0.73	±0.57
Manual Dexterity (T2)	0.97	0.31	-0.17	0.37	±0.45	-0.63	0.73	±0.86
Aiming & Catching (T1)	0.94	0.03*	0.42	0.37	±1.12	-0.35	0.73	±0.48
Aiming & Catching (T2)	0.95	0.10	0.25	0.37	±0.66	-0.62	0.73	±0.84
Balance (T1)	0.92	0.008**	-0.33	0.37	±0.87	-0.63	0.73	±0.86
Balance (T2)	0.91	0.003**	-0.28	0.37	±0.74	-1.20	0.73	±1.64

Note. SE = standard error; Sig. = significance; * = significant at 0.05; ** = significant at 0.001; Statistically significant results indicate that the assumptions have not been met

Component Scores. The analysis of the Shapiro-Wilks test once again showed that three out of the eight scores were significant at $p < 0.05$ or lower, indicating that these domains (total test score (T1), balance (T1), and balance (T2)) are not normally distributed. The remaining five items revealed normal distribution. All domains of the assessment, with the exception of aiming and catching at the first (T1) and second testing (T2) had skewness values that were less than zero, indicating that the data was skewed. The two balance domains, at T1 and T2, had skewness statistics that were less than -1 indicating that they were highly skewed. Examining the standard error values for skewness also supported this pattern for the balance domains. The absolute values for balance (T1 and T2) were twice as high as the standard error, thus indicating the distribution was not symmetric. The TTS and manual dexterity skewness statistics, for both testing occasions, were between -1 and -0.5, and thus were moderately skewed. The aiming and catching domains demonstrated a low skewness and therefore represented a symmetric distribution.

Three out of the eight domains (manual dexterity (T2), aiming and catching (T1 and T2)) had kurtosis statistics that were below zero, indicating a platykurtic distribution of the data. However, as all three of these values were not large this indicated that the distributions were only marginally platykurtic. The remaining five domains had kurtosis statistics above zero, indicating a leptokurtic distribution, meaning that there are more data points in the tails of the distribution than around the mean. The absolute value for the kurtosis of the balance domain (T1) was more than double the standard error (SE = 6.32), revealing a lack of normality based on the kurtosis values (see Table 2). Similar to the standard scores, the kurtosis statistics revealed that the distribution was more platykurtic across all domains on time two compared to time one. The

balance domain showed the largest difference in kurtosis statistics, between time one and time two, with a fluctuation of almost six units.

Table 2.

Results for Tests of Normality, Skewness and Kurtosis for Component Scores Across First (T1) and Second (T2) Testing Sessions.

Component Scores	Normality (Shapiro-Wilks)		Skewness			Kurtosis		
	Statistic	Sig.	Statistic	SE	Normal Distribution	Statistic	SE	Normal Distribution
Total Test Score (T1)	0.95	0.05*	-0.88	0.37	±2.35	1.33	0.73	±1.82
Total Test Score (T2)	0.96	0.13	-0.61	0.37	±1.64	1.17	0.73	±1.60
Manual Dexterity (T1)	0.97	0.26	-0.59	0.37	±1.58	0.16	0.73	±0.22
Manual Dexterity (T2)	0.96	0.18	-0.51	0.37	±1.37	-0.18	0.73	±0.25
Aiming & Catching (T1)	0.96	0.13	0.34	0.37	±0.91	-0.47	0.73	±0.63
Aiming & Catching (T2)	0.96	0.22	0.23	0.37	±0.62	-0.78	0.73	±1.07
Balance (T1)	0.79	0.001**	-2.09	0.37	±5.60	6.32	0.73	±8.62
Balance (T2)	0.84	0.001**	-1.18	0.37	±3.15	0.48	0.73	±0.66

Note. SE = standard error; Sig. = significance; * = significant at 0.05; ** = significant at 0.001; Statistically significant results indicate that the assumptions have not been met

Percentile Scores. Six out of the eight scores, aside from TTS (T2) and aiming and catching (T2), had significance values at $p < 0.05$ for Shapiro-Wilks test (Table 3). This indicates that the data from these areas of the assessment were not normally distributed. The balance scores at time 2 had a significance of $p < 0.001$, demonstrating that it was highly unlikely that the data would be normally distributed. The two domains that did not exhibit statistically significant values were TTS (T2) and aiming and catching (T2).

Across both testing sessions, all but one domain revealed skewness statistics between -0.5 and 0.5, which indicates that the distribution was approximately symmetrical. The balance (T2) domain had a statistic of -0.62, indicating that the distribution was moderately skewed to the left. The standard error for the skewness of the percentile scores was 0.37 across all domains, and all the skewness statistics were within two standard errors. These findings suggest that the data is likely to be symmetric and normally distributed.

All four of the domains, across both testing sessions, revealed kurtosis statistics that were below zero, thus exhibiting a platykurtic distribution. This means that the central peak was lower and broader as compared to the normal distribution. Similarly to the skewness results, all of the kurtosis statistics were less than double the standard error (0.73), indicating an approximate normal distribution. Contrary to the standard and component scores, the percentile scores did not demonstrate the same pattern with the kurtosis statistics across time one and time two. Instead, the kurtosis statistic increased from time one to time two for the total test score and the balance domain, whereas the manual dexterity and aiming and catching domain decreased.

Table 3.

Results for Tests of Normality, Skewness and Kurtosis for Percentile Scores Across First (T1) and Second (T2) Testing Sessions.

Percentile Scores	Normality (Shapiro-Wilks)		Skewness			Kurtosis		
	Statistic	Sig.	Statistic	SE	Normal Distribution	Statistic	SE	Normal Distribution
Total Test Score (T1)	0.94	0.05*	-0.43	0.37	±1.14	-0.81	0.73	±1.11
Total Test Score (T2)	0.95	0.07	-0.35	0.37	±0.93	-0.64	0.73	±0.88
Manual Dexterity (T1)	0.94	0.03*	0.05	0.37	±0.13	-1.19	0.73	±1.62
Manual Dexterity (T2)	0.91	0.003*	-0.20	0.37	±0.53	-1.41	0.73	±1.92
Aiming & Catching (T1)	0.92	0.01*	0.29	0.37	±0.77	-1.11	0.73	±1.51
Aiming & Catching (T2)	0.92	0.10	0.26	0.37	±0.70	-1.15	0.73	±1.57
Balance (T1)	0.90	0.002*	-0.44	0.37	±1.17	-0.88	0.73	±1.21
Balance (T2)	0.89	0.001**	-0.62	0.37	±1.66	-0.70	0.73	±0.95

Note. SE = standard error; Sig. = significance; * = significant at 0.05; ** = significant at 0.001; Statistically significant results indicate that the assumptions have not been met

In summary, the analysis of the normality of the distributions for the TTS and three sub-components revealed that overall the standard scores exhibited the characteristics of normally distributed data most frequently (see Table 4). As a result, the findings based on the component and, in particular, percentile scores should be treated with caution. In terms of the three sub-components, the balance domain violated the most assumptions, as compared to the manual dexterity and aiming and catching sub-sections.

Table 4.

Number of Domains (out of 8) that met the Normality Assumptions Across First (T1) and Second (T2) Testing Sessions.

	Type of Score		
	Standard Scores	Component Scores	Percentile Scores
Normality (Shapiro-Wilks)	5/8	5/8	2/8
Skewness	8/8	2/8	7/8
Kurtosis	8/8	7/8	8/8

Reliability

Test Re-test.

Standard Scores: Total Test Score. Based on the analysis of the standard scores, the hypothesis regarding the TTS, which stated that the ICC would be moderate to high (> 0.70), was not supported. The ICC coefficient was 0.67, indicating that the results only approached the expected degree of reliability. The additional analysis of variance (t-test) also confirmed the inconsistencies between the testing sessions, as scores at time one (M =10.55, SD = 2.49) were significantly lower as compared to time two (M = 11.53, SD = 2.53) ($t(39) = -2.53$, $p < 0.05$).

As evident from the scatterplot, reflecting the correlation coefficient for the TTS (Figure 1, top left), the data set was homoscedastic, but it also demonstrated a restricted range.

Sub-Component Scores. The hypothesis regarding the degree of reliability for the sub-components was not supported, however, the correlation coefficients did approach the expected degree of consistency. The analysis of the manual dexterity sub-component found an ICC = .68, which indicates a moderate degree of consistency. However, the analysis of variance also showed that there was a statistically significant difference between the results from time one (*M = 9.85, SD = 3.10*) to time two (*M = 10.70, SD = 3.32*) (t (39) = -2.14, $p < 0.05$). The scatterplot (Figure 1) for the respective sub-component showed that the correlation for manual dexterity was stronger and had a greater variance around their hypothetical line of best fit, as compared to aiming and catching and balance. In all three scatterplots less than forty data points are visible, even though forty children were examined. This indicates that some children achieved the same scores due to floor or ceiling effects. The scatterplot (Figure 1) showed that the manual dexterity sub-component had a greater variance around the hypothetical line of best fit, in comparison to the other domains.

The aiming and catching sub-component revealed an ICC = .65, and surprisingly no statistically significant differences between times one (*M = 9.95, SD = 2.31*) and time two (*M = 10.13, SD = 2.54*) (t (39) = -0.54, $p = 0.59$) were found. The respective scatterplot (Figure 1, bottom left) showed that the data points were more homogeneous compared to the manual dexterity domain. Also, similarly to aiming and catching, despite a relatively moderate degree of consistency (ICC = .66), the results from t-statistics did not reveal statistically significant differences for the balance sub-component between time one (*M = 11.55, SD = 2.57*) and two (*M = 12.10, SD = 2.18*) (t (39) = -1.79, $p = 0.08$). The bottom right scatterplot (Figure 1)

demonstrates a lack of variability within the data set for this domain. However, the ICC was still approaching a moderate reliability and therefore participants must have scored similarly across trials.

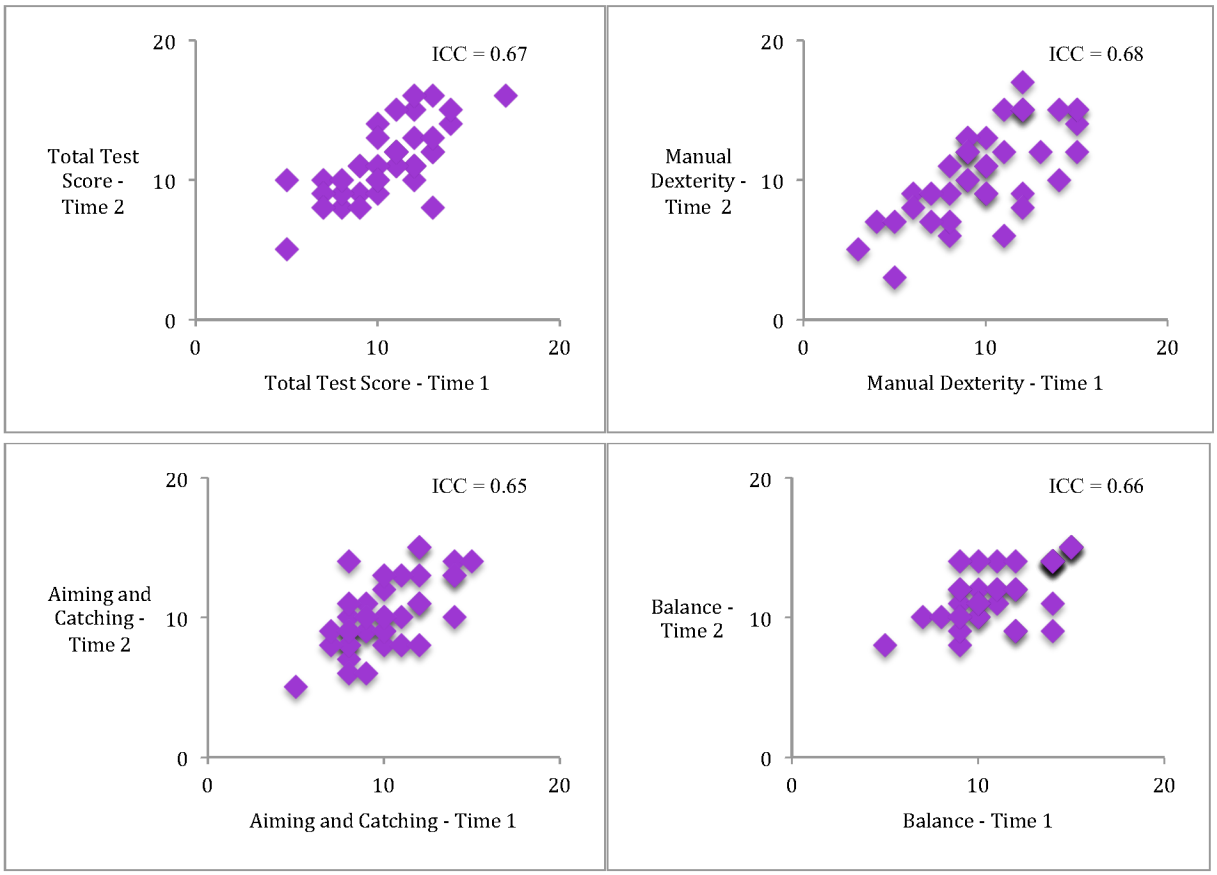


*Figure 1.*Intraclass correlation coefficients and respective scatterplots for standard scores for total test score (top left), manual dexterity (top right), aiming and catching (bottom left), and balance sub-components (bottom right).

Component Scores: Total Test Score. The results indicated that the hypothesis for the TTS was supported, as the ICC coefficient was 0.70, indicating a moderate reliability coefficient. However, the results for the analysis of variance, based on the component scores showed, that on average, the participants achieved a lower TTS on time one (M =80.78, SD = 10.51) as

compared to the results from time two (M = 83.73, SD = 10.34) (t (39) = -2.40, $p < 0.05$). The analysis of the corresponding scatterplot (Figure 2, top left), showed that the data was heteroscedastic, and had a restricted range, but that there were not many outliers. Outliers are operationalized as defined as an observation (data point) that is located substantially away from the line of best fit (Liu & Zumbo, 2007).

Sub-Component Scores. The hypotheses for the sub-component scores were partially supported as the manual dexterity had an ICC value of 0.71, indicating a moderate reliability. However, this was not true for the aiming and catching, and balance sub-components, which did not have a moderate or high reliability. However, in terms of manual dexterity sub-component the analysis of variance showed statistically significant differences between time one (M = 28.50, SD = 6.86) and time two (M = 30.30, SD = 6.79) (t (39) = -2.26, $p < 0.05$). The opposite was true for aiming and catching (t (39) = -0.35, $p = 0.73$), and balance sub-components, which revealed no statistically significant differences between the means (t (39) = -1.87, $p = 0.07$). The analysis of the scatterplot for manual dexterity showed a substantial variance around the hypothetical line of best fit. On the other hand, the aiming and catching domain was characterized by a restricted range, as the data was not systematically spread along the hypothetical line of best fit. Lastly, the balance sub-component closely resembled the data pertaining to TTS, as a restricted range as well as heteroscedasticity was evident.

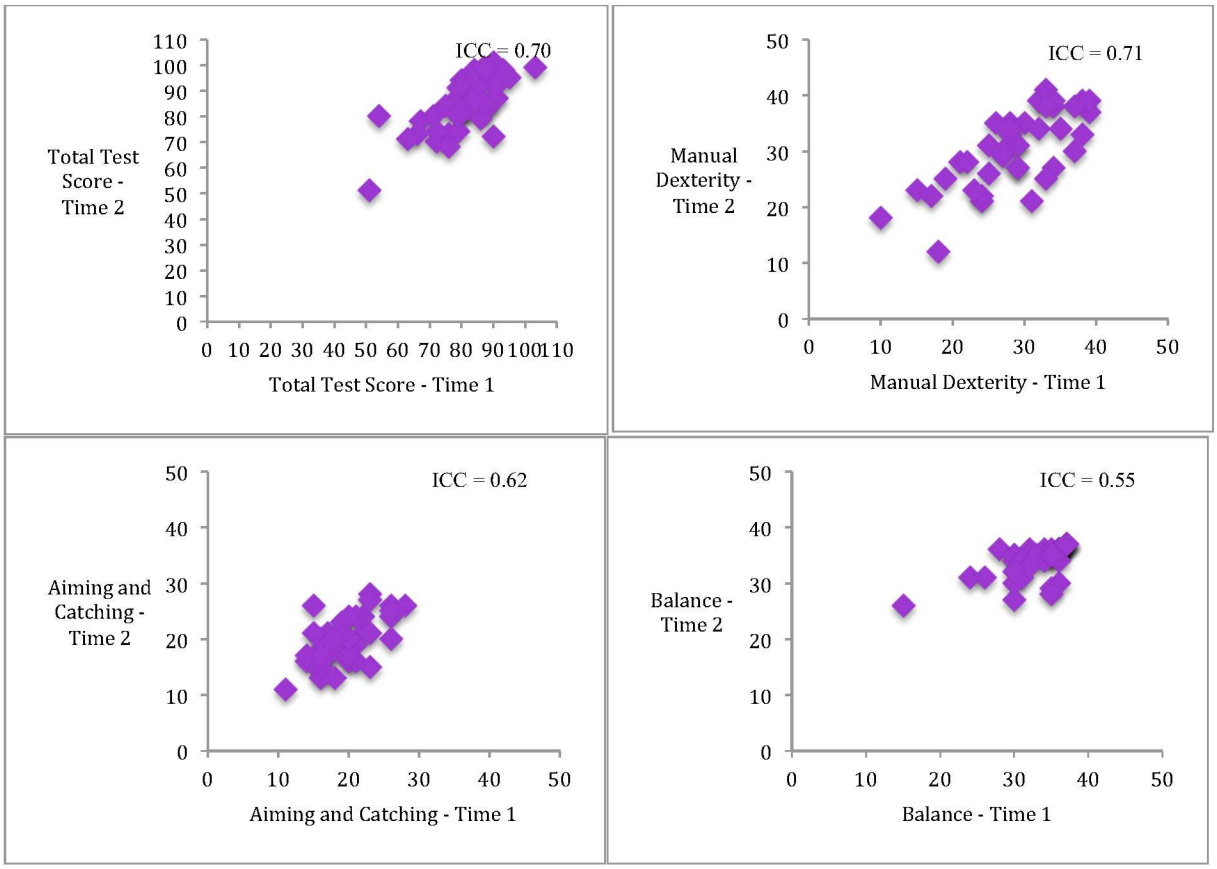


Figure 2. Intraclass correlation coefficients and respective scatterplots for component scores for total test score (top left), manual dexterity (top right), aiming and catching (bottom left), and balance sub-components (bottom right).

Percentile Scores: Total Test Score. The hypothesis was not confirmed for the TTS, however the correlation approached the expected value with an ICC of 0.68. Nevertheless, the additional analysis of variance indicated that the percentile scores, on average, were lower on time one (M =56.20, SD = 25.64) as compared to the results from time two (M = 63.00, SD = 24.47) ($t(39) = -2.21$, $p < 0.05$). As evident from the corresponding scatterplot (Figure 3), the TTS demonstrated a relatively equal variance around the hypothetical line of best fit, and therefore the data was homoscedastic. As well, there were no evident outliers.

Sub-Component Scores. In terms of percentile scores, the hypothesis was not supported, as none of the sub-components of the assessment had an ICC > 0.70. The ICC values were 0.64, 0.63, and 0.42 for manual dexterity, aiming and catching, and balance, respectively. Nevertheless, the manual dexterity sub-component showed no statistically significant differences between time one ($M = 48.93$, $SD = 56.08$) and time two ($M = 56.08$, $SD = 31.76$) ($t(39) = -1.76$, $p = 0.09$). Similarly, aiming and catching also revealed no differences between time one ($M = 49.28$, $SD = 25.34$) and time two ($M = 50.05$, $SD = 27.17$) ($t(39) = -0.22$, $p = 0.83$). But, the analysis of variance of the balance sub-component showed statistically significant differences between time one ($M = 65.03$, $SD = 25.64$) and time two ($M = 67.88$, $SD = 24.18$) ($t(39) = -0.67$, $p < 0.50$). The analysis of the scatterplots showed that the data points in all three sub-component plots were not close to the hypothetical line of best fit, thus indicating heteroscedasticity. Similarly to the other types of scores, forty data points were not evident, therefore decreasing the amount of variance within the data set.

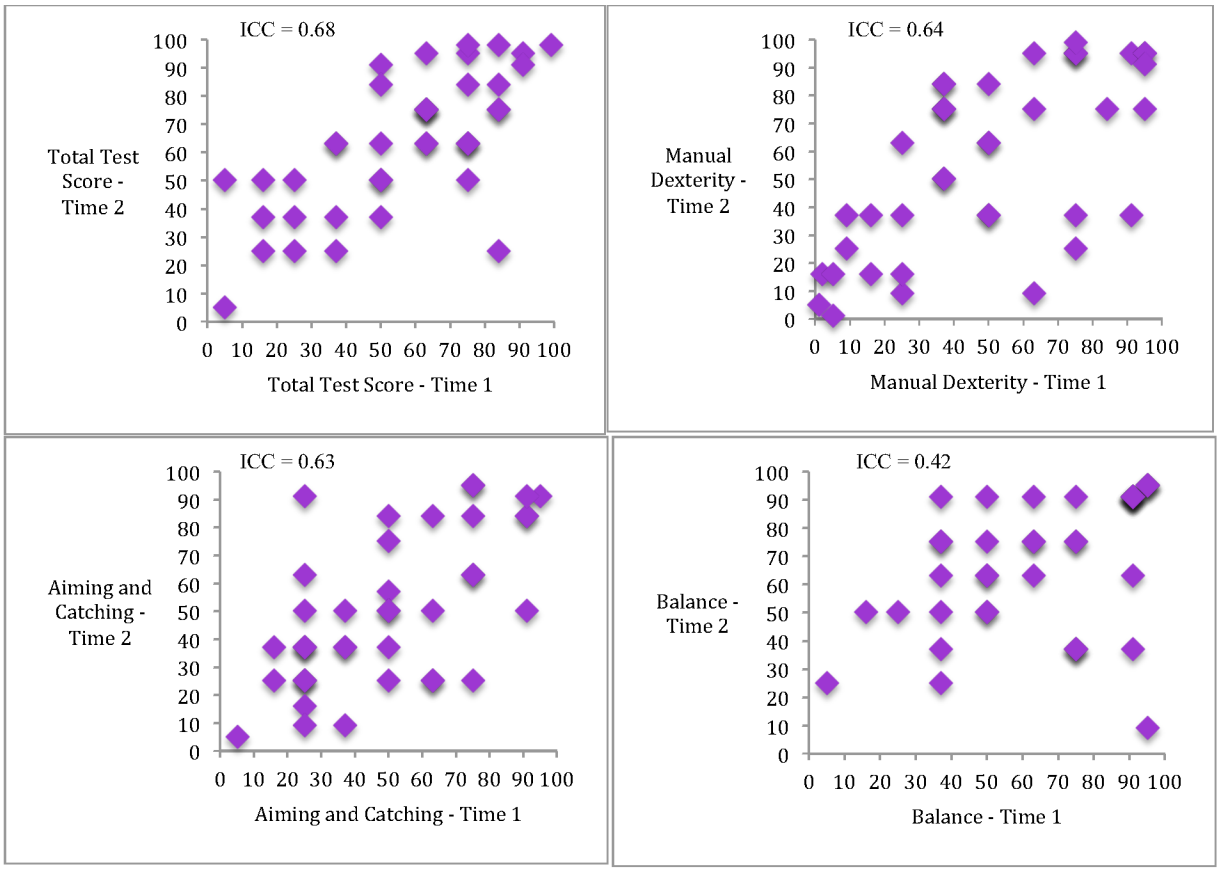


Figure 3. Intraclass correlation coefficients and respective scatterplots for percentile scores for total test score (top left), manual dexterity (top right), aiming and catching (bottom left), and balance sub-components (bottom right).

Internal Consistency. There were two different testing sessions (T1 and T2) that could be used for the analysis of internal consistency, and for the purpose of this study data from time one was analyzed. The analysis of normality assumptions provided similar results for both times, therefore time one was chosen to ensure that the memory and/or practice effect did not affect the results. The analysis was carried out on the item standard scores, which were derived from the raw scores from the eight items of the MABC-2. These scores were used to examine the internal consistency of each of the three sub-components; manual dexterity, aiming and catching, and balance.

Manual Dexterity. The results did not support the hypothesis, as Cronbach's alpha was 0.61, representing only acceptable internal consistency for this sub-component. Further analysis was implemented using Cronbach's alpha with items deleted to determine which, if any, specific items affected the low scores. The analysis showed that the coefficient did not increase when any of the three items were deleted, indicating that one specific task did not jeopardize the internal consistency of this set of items (see Table 5).

Aiming and Catching. Based on previous results, it was hypothesized that this sub-component would have the highest internal consistency, compared to manual dexterity and balance. The data did not support this hypothesis as the aiming and catching domain had a Cronbach's alpha of 0.49 indicating a questionable internal consistency. Cronbach's alpha with items deleted could not be calculated for the aiming and catching sub-component, as there are only two tasks within this sub-section.

Balance. As it was the case with the two previous sub-sections, the findings from the present study did not support the hypothesis for balance, as the Cronbach's alpha was 0.53. When items were deleted from this sub-component (Table 5), the Cronbach's alpha value did not increase for any of the three items, indicating that no one specific item caused the questionable internal consistency.

Table 5.

Cronbach's Alpha with Items Deleted for the Item Standard Scores for Time One.

Item	Cronbach's Alpha if Deleted
Manual Dexterity – Cronbach's Alpha = 0.61	
Manual Dexterity 1	0.54
Manual Dexterity 2	0.54
Manual Dexterity 3	0.41
Aiming and Catching – Cronbach's Alpha = 0.49	
Aiming and Catching 1	-
Aiming and Catching 2	-
Balance – Cronbach's Alpha = 0.53	
Balance 1	0.51
Balance 2	0.41
Balance 3	0.41

Note. There are no values for aiming and catching with items deleted, as an item cannot be deleted since there are only two items in this domain.

Standard Error of Measurement.

Total Test Score. To infer the degree of absolute reliability the analysis of the SEM was implemented. The SEM for the TTS, based on the standard scores, revealed a value of ± 1.80. This indicated that although there is measurement error present in the TTS data, it is small enough in this context that it does not have a large impact on the spread of scores around the mean (10.55). Analysis of the component scores revealed a larger SEM of ± 7.39. The percentile scores for the TTS revealed the largest SEM with a value of ± 18.59. This value indicated that

the spread of scores around the mean of the observed score would be very large and there would be a high degree of variability.

Sub-Component Scores. The manual dexterity sub-component for the standard scores revealed a SEM of ± 2.25. The SEM for the component and percentile scores were ± 4.76 and ± 22.97. Analysis of the aiming and catching sub-component showed that the SEM scores were ± 1.78, ± 3.16, and ± 20.03 for the standard, component, and percentile scores, respectively. Similarly to the other domains, the SEM of the balance domain for the standard scores was ± 1.92, and it increased to ± 3.49 and ± 23.55 for the component and percentile scores.

Discussion

The aim of this study was to examine the selected aspects of the reliability of MABC-2 for age band 2 (7 to 10 years old), for the standard, component, and percentile scores. In the MABC-2 manual, the authors have included data regarding the reliability of this assessment tool. However, most of this research was based on the original assessment test (MABC, 1998) assuming that the second version would reveal similar results. Little research, in regards to reliability, has previously been conducted on this specific age band and only a small amount of information has been collected about the reliability of age band 1 and 3. In this study, test-retest approach was used to examine the stability of the total test score and the 3 sub-sections of MABC-2. Internal consistency examined the reliability of scores in the sub-components, and the SEM was used as an absolute estimate of reliability.

Tests of Normality

The results from the tests of normality across all three types of scores were mixed and somewhat equivocal. As evident from the summary table (Table 4), the results indicated that the standard scores exhibited the desired characteristics of normally distributed data. All 8 of the

areas demonstrated minimal skewness and kurtosis, and 5 out of the 8 areas were normally distributed, according to the Shapiro-Wilks test. Also as evident (Table 4), the component scores demonstrated skewness for 6 out of the 8 areas, thus indicating that these sub-components were not normally distributed. The percentile scores met the normality assumptions based on the skewness and kurtosis tests; however only 2 out of the 8 areas were normally distributed according to the results from the Shapiro-Wilks test. When the data is not normally distributed, the inferences need to be treated with caution.

In terms of homo/heteroscedasticity, when the data is heteroscedastic "individuals who score the highest values on a test also show the greatest amount of measurement error … and smallest changes in responses" (Atkinson & Nevill, 1998). Therefore it may be difficult to identify these small changes in participants that are performing the best, even though these changes might allow for the detection of measurement error. Also, it is interesting to note that in many instances the values of the ICC coefficient indicated a moderate degree of consistency, yet the analyses of variance showed statistical differences between the two testing sessions, and vice-versa, no differences were found despite low ICC values. This indicates that although both analyses are expected to reveal similar results, they are confounded by different factors. Often when the correlations are high, a certain level of measurement error can be accepted (Atkinson & Nevill, 1998). As well, the results of the paired t-test may conclude that there were no statistically significant differences between the groups, when at individual level there may be substantially differences, especially when the sample is heterogeneous (Atkinson &Nevill, 1998). Due to this fact, if the results are contradictory, the scenario with the lower reliability coefficient should be trusted.

Test Re-test

Standard Scores.

Total Test Score. The analysis of the TTS revealed an ICC of 0.67 when test re-test reliability was examined, thus indicating a moderate degree of reliability. The analysis of the scatterplots showed that the weak relationship between the first and second testing session could be due to the restricted range of the data, which may have caused the data to be homoscedastic. As the use of correlations alone can be problematic to assess the degree of systematic bias, particularly when the inter-individual variability is low, the analysis of variance was also carried out. In line with the correlation, the data showed significant differences between time one and time two. Looking at the individual data for the TTS, only 9 out of the 40 participants achieved the same standard score across the two trials. This indicated that there is a lack of consistency across performances in more than three quarters of the individuals. Of the 31 participants scoring differently across trials, 22 actually scored higher on the second trial potentially due to a learning effect. As a result, the lack of consistency between the two testing times should not only be attributed to the amount of variability present, but also due to systematic differences between the two testing conditions, as confirmed by the t-tests.

To date, there have been no other studies conducted on age band 2 of MABC-2 that examined the reliability of the standard scores. In relation to the other age bands from MABC-2 (age band 1 and 3), there were two studies conducted that examined the test re-test reliability based on the standard scores (Ellinoudis et al., 2011; Wuang et al., 2012). The ICC values from the present study were much lower when compared to results by Ellinoudis and colleagues (2011) (ICC = 0.85) and Wuang and colleagues (2012) (ICC = .97) for the TTS of age band 1. The discrepancies could be attributed to many factors, aside from the plausibility that this age

band is less reliable as compared to the others. The former study incorporated 183 children, whereas the study by Wuang et al., (2012) involved a sample of 144 atypically functioning individuals. In both cases the inter-individual variability present in the sample may have contributed to larger ICC values. The authors of both studies did not report on actual differences in means between the two testing conditions. In terms of the research examining the test re-test reliability based on the previous version of MABC, Chow and Henderson (2003) found that the TTS had a moderate reliability (0.77). The reliability coefficient that emerged from that study may be higher when compared to the present value due to the increased sample size (75 participants). The study was also conducted on age band 1 (4 to 6 years), and developmentally there may be more variability within this age band as compared to the older ages. The authors did not provide the type of score that was used (standard, component, or percentile) for the calculation of ICC. As a result, caution is warranted when comparing the results from the two studies.

Sub-Component Scores. The analyses of the stability for manual dexterity revealed the ICC of 0.68, which indicated that the coefficient was approaching a moderate reliability. Surprisingly, the analysis of the scatterplot (Figure 1) showed that the data appeared to be normally distributed around the hypothetical line of best fit, suggesting that factors such as restricted range or outliers are not responsible for the low correlation value. The lack of consistency was also supported by the analysis of variance, which showed that there were statistically significant differences between the two testing sessions. Looking at the individual data across time one and time two, 25 of the 40 participants scored higher on time two suggesting that some systematic and/or random bias emerged. Of the remaining fifteen participants, only three individuals had the same score across both testing sessions (Appendix

G). The standard deviation also increased from 3.10 at time one to 3.32 at time two, which showed that there was more variance within the data set from the second testing. In comparison to previous research involving the MABC-2, all of the studies conducted reported a higher ICC for manual dexterity as compared to those found here. Ellinoudis et al., (2011) revealed an ICC of 0.82 for the manual dexterity of age band 1. Wuang and colleagues (2012) reported an ICC of 0.97 for the test re-test reliability, which represented the highest ICC value of all of the existing studies. This study had a sample that consisted of a larger age range (6 to 12 years), thus creating a larger variance thus possibly artificially inflating the reliability coefficient. No studies were conducted on age band 3, or the original MABC, that examined the test re-test reliability for manual dexterity.

In terms of aiming and catching, the ICC for the test re-test reliability was 0.65, indicating a questionable consistency. An analysis of variance showed that there were no statistically significant differences between time and time two, therefore it is expected that the reliability coefficient would be higher. The individual data (Appendix G) showed that only 9 of the 40 participants had the same standard score across both testing occasions. Therefore, the ICC may be low due to the lack of reliability from the participants scoring different across the two trials. The analysis of the scatterplot showed that the low degree of reliability could also be due to the restricted range of the data, as perhaps the tasks were too easy for the individuals resulting in a ceiling effect. In regards to previous research, similar results (ICC = .61) were reported by Ellinoudis and colleagues (2011), despite the fact that their sample was much larger than the current data set. This value was much lower than the ICC findings for the TTS and the other sub-components in the same study. These findings were similar to those from the present study, where the aiming and catching sub-component had the lowest coefficient (ICC = .65) for all

other domains. One potential reason for the lower reliability coefficient may be the fact that the aiming and catching is comprised of only two tasks (two handed catch and throwing a bean bag on to a target). This lower number of items could have affected the reliability coefficient. In contrast, Wuang et al., (2012) found that the aiming and catching section of the assessment had a substantially higher ICC for test re-test (0.91). However, as previously mentioned, the study conducted by Wuang and colleagues encompassed a much wider age range with varying ability levels. Thus, the addition of different age bands would add a greater variance to the results, therefore potentially inflating the ICC value.

In terms of the balance sub-component, the data showed an ICC of 0.66, based on the standard scores. The reliability coefficient indicated that this section approached moderate degree of reliability. Analysis of the scatterplot (Figure 1, bottom right) demonstrated a restricted range, therefore decreasing the variance and subsequently the value of reliability coefficient. The restricted range may have been caused by one of the tasks being too difficult for the participants. For example, most of the individuals found the one-board balance task to be very challenging and almost all scores coincided with poor performance. The scatterplot also showed that there were less than forty data points, which supports the previous point. In addition, the analysis of variance revealed no significant differences between time one and time two. Once again, this indicates that the individuals were scoring consistently poorly across both sessions, but the low ICC may be potentially a result of restricted range. In fact, in the balance sub-component 50% of participants achieved the same score across trials, in comparison to the other two sub-components and the TTS.

In regards to previous research, all studies reported higher ICC values for the balance sub-section as compared to the present results (Ellinoudis et al., 2011; Wuang et al., 2012).

Ellinoudis and colleagues (2011) reported an ICC value of 0.75 for age band 1. However, it should be noted that the nature of the tasks is different across the age bands. For example, in age band 1 the participants are required to complete 3 tasks including one leg balance, walking with heels raised and jumping on mats. Although the latter two items are comparable to those in age band 2, the one leg balance task would be considerably easier than the one-board balance task. Adding the balance board further contributed to the level of difficulty as most children exhibited poor performances. Therefore, this task likely contributed to a restricted range in age band 2. Furthermore, the difference in the strength of the reliability coefficient, as compared to the present study, could be due to the size and nature of the sample. Ellinoudis and colleagues assessed 183 children, between 3 and 6 years of age, which is a substantially larger and younger sample as compared to the one used in this study. Also, Wuang and colleagues (2012) reported an ICC of 0.97, while testing atypically functioning individuals. Thus, a larger and more heterogeneous sample size likely contributes to stronger reliability coefficients, regardless if the tasks/or performances are actually more stable/reliable.

Component Scores.

Total Test Score. The analysis of the TTS for the component scores revealed an ICC of 0.70, indicating a moderate reliability. These findings were not supported by the analysis of variance, which showed that there were significant differences between the results from time one to time two. Of the forty participants, only two achieved the same score across the trials (Appendix H). Furthermore, ten out of the remaining thirty eight individuals scored within two points of their first testing session. Therefore, it can be concluded that although the participants scored differently on time two, the overall rankings did not change and that is why the ICC was

0.70. The scatterplot also showed that the moderate degree of reliability could be due to a substantially restricted range (Figure 2, top left).

In regards to previous research, only one other study has been conducted on age band 2 of the MABC-2. Holm and colleagues (2013) examined intra- and inter-rater reliability based on component scores and reported ICC values of 0.68 and 0.62, respectively. These findings are similar to those reported here. The sample size and characteristics of the participants were also comparable between the studies. Thus, thus the emerging results appear to be robust even though different reliability coefficients were examined. To date, more research has been conducted on the other age bands using the component scores. Smits-Engelsman and colleagues (2011) reported an ICC of 0.83 for the TTS of age band 1. The methodology of the previous study was comparable to the present study, as the children were assessed one to two weeks apart, and the test re-test reliability was measured using ICC. The sample size and characteristics were also comparable as the present study had 40 participants whereas the previous research by Smits-Engelsman and colleagues (2011) involved 50 typically functioning children. Thus it appears that age band 1 may be more reliable than age band 2, at least in regards to the component scores. In fact, the authors of MABC-2 changed three of the tasks for age band 1, compared to the original tool, whereas two previous age bands were combined to form age band 2 and therefore more of the tasks were changed or had revisions.

Sub-Component Scores. The analysis of the manual dexterity sub-component revealed an ICC of 0.71, representing a moderate degree of reliability. The analysis of variance showed that there were statistically significant differences between time one and time two for this sub-component. Examination of the individual data showed that out of the forty participants, only two of them achieved the same score across trials and eleven of the children were within two

points of their initial score (Appendix H). Twenty-seven participants scored substantially higher or lower across trials, thus revealing lack of stability of the scores. Further analysis of the scatterplots indicated that there was an even distribution of the data around the higher parts of the hypothetical line of best fit. However, a restricted range was also evident as there were few data points in the lower quadrant of the plot, indicating that the tasks within the manual dexterity sub-component were too difficult creating a floor effect. Likely, this deflated the ICC value due to the lack of variance within the participants' scores.

In terms of previous research, the present results are higher than those reported in the past studies. Holm and colleagues (2013) showed that age band 2 had ICC values of 0.62 (intra-rater) and 0.63 (inter-rater) for the manual dexterity sub-component. In relation to the other age bands of MABC-2, Smits-Engelsman and colleagues (2011) reported an ICC of 0.85. This value was the highest ICC value reported for manual dexterity across all studies, based on the component scores. The ICC might be higher than the one found in the present study because Smits-Engelsman and colleagues examined age band 1. Hence, younger children may show more variance within their scores, therefore increasing the reliability coefficient.

The analysis of the aiming and catching sub-component, based on the component scores, revealed an ICC of 0.62, which represents a questionable reliability. This was not supported by the analysis of variance, as there were no statistically significant differences between time one and time two of the assessment. However, analysis of the scatterplot did reveal that the data had a restricted range likely because the two tasks in this section were too easy for the individuals. Most of the participants were able to complete the tasks of catching with two hands and throwing bean bags on to a mat with at least a 50% success rate. As a result, there were no scores in the lower percentiles resulting in homoscedastic data set. Thus, although there were no statistically

significant differences across the trials, due to the restricted range and lack of variance the reliability coefficient was still low. In terms of the previous research, Holm and colleagues (2013) also examined the component scores and reported ICC values of 0.49 and 0.77 for intra- and inter-rater reliability, respectively. Smits-Engelsman and colleagues (2011) used the same scores to test the reliability of the aiming and catching sub-section from age band 1. The authors reported an ICC of 0.74, which is moderate but still higher than the one found here. Once again, this reliability coefficient may be higher than the one found here due to the fact that different age band was examined, which may be more reliable. Another reason for the emerging differences may be that the younger individuals, although also typically functioning, may be more variable as compared to those who are 7 to 10 years of age.

The analysis of the balance sub-section revealed an IC of 0.49, which demonstrates a weak reliability. This finding was not supported by the analysis of variance, as no significant differences were found between time one and time two. This means that the children's results and rankings could have changed from time one to time two, demonstrating a lack of stability, while the means remained the same. The analysis of the scatterplot provided additional support for the weak reliability. Almost all of the data points were clustered together in a small area indicating restricted range. This lack of range could be due to difficulty of the tasks embedded in this section. The low reliability found here is consistent with the coefficients (.49 & .29) reported in previous research, for age band 2. In fact the study by Holm and colleagues used a similar sample in terms of its size, thus the pattern appears to be robust. The authors suggested that the low ICC values may be due to the one board balance task as it had a high SEM and therefore lacked consistency (Holm et al., 2013). This is in line with present study which also revealed that one board balance task had a high SEM. In some cases, participants would achieve

a perfect score (30 seconds) on time one, however during time two they would only be able to balance for 5 seconds. This task proved to be inconsistent from trial to trial, within each testing session, as well as between the two testing sessions.

Percentile Scores.

Total Test Score. The analysis of the TTS based on the percentile scores revealed an ICC of 0.68, indicating that the reliability is approaching a moderate strength. This was supported by the analysis of variance, which showed that there were significant differences between the group means when time one and two were compared. The individual data (Appendix I) showed that only nine of the forty participants exhibited same/similar percentile scores across trials. Among the remaining individuals the performances improved on the second testing session, as shown by a higher percentile. Based on the analysis of the scatterplots, it was evident that there were absolute differences across the individuals when both testing sessions are compared. There were also a few outliers as shown in the data (Figure 3, top left). As a result, the standard deviations for both time one (25.64) and time two (24.47) were equally as high. Overall, the results showed a robust pattern of lack of consistency among the group, and across the trials.

In regards to previous research, there have been no studies conducted on the reliability of the assessment based on the percentile scores for the MABC-2, for any of the three age bands. Croce and colleagues (2001) examined the reliability of the original version of the MABC using the percentile scores. Croce et al. (2001) did not report a reliability coefficient for the TTS, however each of the ten individual items had ICC values ranging from 0.92 to 0.98. From these results, one can infer that the reliability of the TTS would have been high. The percentile scores are the most commonly used scores by clinicians, as they classify individuals into different categories. Therefore, it is interesting that no studies have examined the reliability of the

assessment tool based on these scores since they are the scores that are being used in clinical settings. The results from the present study showed that the distribution of the percentile scores was jeopardized to the largest extend, and they also exhibited the highest SEM due to low reliability and high variability. Thus, due to this constraints caution is warranted when interpreting these scores as they may change if the person/child was to be retested again

Sub-Component Scores. The analysis of the manual dexterity sub-component, based on the percentile scores, revealed a questionable reliability of 0.64. This finding was supported by the analysis of variance, as there were significant differences found between the group means of time one and time two. Furthermore, the scatterplot showed that the questionable degree of reliability could be due to the outliers and restricted range in the data as there were fewer than forty data points evident in the plot. This indicates that some individuals achieved the same percentile scores across testing sessions, likely due to floor effect as some tasks were too difficult. The aiming and catching sub-component had a reliability of 0.63, indicating a questionable reliability. The analysis of variance showed that there were statistically significant differences between time one and time two for this sub-component, which supports the ICC findings. Only two individuals had the same scores across the testing sessions (Appendix I), and of the 38 participants whose scores fluctuated, 25 achieved a better score on the second testing session. This fact may be potentially attributed to the learning effect. The balance sub-component for the percentile scores revealed an ICC of 0.42, once again indicating a weak reliability. This was confirmed by the analysis of variance. The overview of the scatterplot also showed that the data was very heteroscedastic (Figure 3, bottom right). Examination of the individual scores indicated that only 16 of the 40 participants remained in the same percentile across time one and time two.

Internal Consistency

Item Standard Scores.

Manual Dexterity. The Cronbach's alpha for manual dexterity from the present study was 0.61 indicating a questionable reliability. As there are only three tasks within the manual dexterity sub-component, this could have contributed to the lower internal consistency, as Cronbach's alpha is higher when there are more test items. Furthermore, an analysis was performed examining Cronbach's alpha with items deleted to infer whether a specific task in the sub-component had an effect on the low internal consistency. The value for Cronbach's alpha did not increase when any of the three manual dexterity tasks were deleted, indicating that one specific task did not jeopardize the internal consistency of this sub-component. The low internal consistency could be due to the fact that one of the three items is not measuring the construct of interest, therefore lowering alpha. For example, the placing pegs and threading lace task are very similar in that they are both timed tasks. However, the drawing trail-2 task is self-paced and it requires effective use of a pen/pencil. Examination of the individual data (Appendix J) showed that thirty-one of the forty participants scored lower on the drawing trail-2 task compared to the placing pegs and threading lace tasks.

In regards to previous research, to date there have been no studies that examined the internal consistency of age band 2 of MABC-2. There has been only one study that examined the internal consistency of individuals across this age group (7-10 years old), however it also included children who were below and above this age band and therefore direct comparisons cannot be made (Wuang et al., 2012). In relation to other age bands, one study examined the internal consistency of age band 1 (Ellinoudis et al., 2011). The study revealed a low internal consistency for the manual dexterity sub-component, with an alpha of 0.51. Wuang et al., (2012)

also reported that the manual dexterity sub-section had the lowest Cronbach's alpha (0.81) in comparison to the other sub-components, even though the magnitude of the correlation was relatively high. It should be noted, however, that the participants in this study were atypically functioning and the age range was relatively large (6 to 12 years of age), which may have created a higher variance thus increasing the reliability coefficient. The manual dexterity sub-component has three tasks within the domain, and the number of items directly affects the internal consistency. The study conducted by Wuang and colleagues (2012) analyzed a larger number of items for each sub-component since the study encompassed three different age bands. Thus, the manual dexterity sub-component would have consisted of 9 items (3 items from age band 1, 2, and 3), as compared to the 3 included in the age band 2.

Aiming and Catching. The present study revealed a Cronbach's alpha of 0.49 for the aiming and catching sub-component. This value represents a low internal consistency for this sub-component. The tasks for this sub-component consist of catching with two hands and throwing a bean bag on to a mat. The first task involves interceptive skills whereas the second one is more off an accuracy task, without imposing external time demands on an individual. With such a low internal consistency, it can be concluded that although both involve goal-directed manual actions they do not belong to the same domain such as ball skills. Also, as Cronbach's alpha is affected by the number of items, the lower consistency may be due to the fact that there are only two items in this sub-section. Out of the two tasks, twenty two of the individuals scored the same or better on the bean bag task, compared to the catching a ball task based on the scores from time one (Appendix J). Since the children in this study were typically functioning, they should have been achieving near perfect scores on all tasks, across all trials, however this was not the case.

In regards to previous research, Ellinoudis and colleagues (2011) found the internal consistency of age band 1 to be acceptable with an alpha of 0.70. The only other study that reported an internal consistency for aiming and catching had an alpha value of 0.84 (Wuang et al., 2012). Thus, three very different results were reported for the internal consistency of the aiming and catching sub-component, and although the values cannot be directly compared due to varying populations, age bands, and scores used for calculation, it is a concern that the internal consistency can have such a large range (0.49 to 0.84).

Balance. The analysis of the internal consistency for the balance sub-component revealed a Cronbach's alpha of 0.53. This low internal consistency may indicate that the three items within this domain may not be measuring the same domain. When examining the analysis involving Cronbach's alpha with items deleted it was evident that when each task was individually removed, the alpha decreased. The walking on a line and hopping on mats tasks, although also completed on one foot, saw much higher scores than the one-board balance task. From the motor control perspective, when performing the first two items children were able to use their sensory input to see control their actions. In the one-board balance task, the participants' visual and proprioceptive sensory input was compromised as the board was placed directly under their foot. This factor could explain the lower scores on this task as compared to the other two. In fact, the individual data revealed that the one board balance task was the most difficult, while the remaining two (walking on a line and hopping on mats) proved to be too easy as almost all children had a perfect score. In terms of the previous research, Ellinoudis and colleagues (2011) also reported a relatively low Cronbach's alpha (0.66), for the balance sub-component of age band 1. In contrast, Wuang et al., (2012) reported the highest internal consistency for balance with an alpha of 0.84. However, as previously mentioned, the findings

from this study may be substantially higher as compared to other studies, because of the larger sample size and the increased number of items due to the span across three age bands. As a result, it appears that reliability of the balance subcomponent is questionable when individual age groups are examined.

Standard Error of Measurement

Standard error of measurement (SEM) is a reliability coefficient, which examines the dispersion of measurement errors for an individual, if he/she were to be tested repeatedly. SEM is different than the previous two coefficients (ICC and Cronbach's alpha), as it is calculated using the reliability coefficients and the standard deviations of the sample. The SEM should only be applied when the within group SD is relatively low. The lower the reliability coefficient, and/or the higher the within-group variance, the larger the SEM will be.

Total Test Score. The SEM based on the TTS produced a variety of results, depending on the type of score being examined (e.g., standard, component and percentile). Among those only one out of the three had a SEM that was acceptable for an assessment tool such as the MABC-2. The standard score revealed a SEM of ± 1.80 for the TTS, which is generally considered to be a moderate SEM, based on the results from previous literature on the MABC-2 (Holm et al., 2013). The component and percentile scores both had high SEM values of ± 7.39 and ± 18.59, respectively. These two types of scores had similar reliability coefficients to that coinciding with the standard scores, however the standard deviation of the sample was much higher for both of them. The component scores revealed a SD of 10.51, while the percentile scores had an even higher SD of 25.64. Ideally, the reliability coefficient would be high and the SD would be low, neither of which is true for this domain. The high SEM for the percentile scores is most concerning from the clinical perspective as this is most often score used to assess

children's performance for the purpose of screening and placement, as well when assessing the effectiveness of different intervention approaches. Thus, the clinicians should use other sources of data regarding the movement proficiency of the child, or if possible, to test the child more than once given that the learning effect can be minimized. If not, the resulting inferences may lead to false positive or negatives inferences, which have equally damaging consequences in regards to the child's psychological and/or physical well-being. As there has been no other study examining the SEM for TTS, based on percentile scores, this issue warrants further verification.

In relation to previous research, on age band 2, one other study had examined the SEM. Holm and colleagues (2013) implemented component scores and reported the SEM values of ± 4.9 and ± 6.8 for intra- and inter-rater reliability, respectively. These values, although still relatively high, were substantially lower as compared to the present data despite the fact that similar samples were used. Thus, it appears that the absolute reliability, as measured by SEM, is questionable for TTS when component scores are implemented. When examining the SEM across other age-bands, and different types of scores, the emerging results were comparable to the present research in some instances but not others. For example, when the SEM of age-band 1 was examined using standard scores, Smits-Engelsman and colleagues (2011) reported relatively low values of 1.24 and 1.37, respectively. This is also in line with the results reported by Wuang and colleagues (2012) who found the low SEM values for the TTS using standard scores. The values reported in these investigations are in the ballpark of the present results. This provides initial evidence that standard scores, regardless of the age-band examined, may be most reliable when SEM for the TTS is examined.

Sub-Components.

Manual Dexterity. The SEM for the manual dexterity, based on the standard scores, was moderate (± 2.25). This value resulted from reliability coefficient which was not high, but acceptable (± 0.68), and a group SD of 3.10. In terms of the component scores, the analysis revealed a SEM of ± 4.76. There was a moderate reliability for the manual dexterity sub-component, based on the ICC value of 0.71. The SD for this domain was 6.86 which would have also led to a larger SEM. Lastly, the analysis of the manual dexterity sub-component, based on the percentile scores, revealed the highest value of ± 22.97. The pattern of results emerging for this subcomponent closely resembles the previously discussed analysis of different type of scores for TTS. Once again, the standard scores were most reliable whereas the percentile values exhibited the most measurement error. In regards to previous research, Holm and colleagues (2013) reported a SEM of ± 3.20 for both intra- and inter-rater reliability for component scores, which is consistent with the present results. This is despite the fact that the reliability for the previous study was questionable (0.62 for intra-rater and 0.63 for inter-rater), whereas the ICC that emerged in the present data was moderate (0.71). The other study, which examined the SEM for this sub-component, reported a value of 0.31 for the standard scores (Wuang et al., 2012). This is substantially lower when compared to the present value. However, as previously mentioned the study conducted by Wuang and colleagues (2012) had a much larger sample size, and the analysis was collapsed across all different age groups which likely artificially inflated the reliability coefficient.

Aiming and Catching. The analysis of the aiming and catching sub-component once again followed the pattern evident in the previous analyses. The SEM for standard scores was the lowest (± 1.78), which resulted from a relatively low reliability coefficient (0.65), and a

moderate SD (2.31). The component scores revealed a slightly higher SEM (± 3.16), as compared to the standard scores, but were substantially lower in contrast to percentile scores (± 20.03). The present results, in regards to the standard and component scores, are in line with previous work (Holm et al., 2013; Wuang et al., 2013), which also revealed acceptable levels of reliability. As no other research involved the analysis of the percentile scores, the validity of inferences pertaining to these scores remained equivocal. Nevertheless, the fact remains that percentile scores exhibited the least amount of reliability across different facets of the test.

Balance. The third sub-component of MABC-2 that was examined had a SEM of ± 1.92 based on the standard scores, ± 3.49 in relation to the component score, and ± 23.55 for the percentile scores. This sub-section had a very weak reliability, with an ICC of 0.42, and the intra-group variability was also consistently high across all the different scores (e.g., SD = 25.64 for percentile scores). Previous research examining the reliability of balance domain reported a SEM of ± 2.7 and ± 4.5 for intra- and inter-rater reliability, respectively (Holm et al., 2013). These findings were similar to those from the present study, reporting moderate to high SEM. Overall, it appears that the SEM fluctuated depending on the type of score that was being examined. The standard scores revealed the lowest SEM values, which was likely due to the low SD values as well as moderate reliability coefficients. The component scores also had moderate reliability coefficient, however the SD of the sample was much higher and therefore increased the SEM. Lastly, the percentile scores exhibited the highest SEM, which was likely due to poor reliability coefficients, as well as the high SD values. All in all, it appears that in terms of SEM the degree of observed reliability was specific to a particular type of score used. This is an important finding for clinical practitioners who should be aware which scores are most reliable and which warrant caution.

General Discussion

The original MABC has been one of the most widely used assessment tools for children with movement difficulties, and it is hoped that its revised version can take on the same gold-standard status. However, due to extensive changes to the test it is necessary to revaluate its basic psychometric properties, particularly for those age bands that have not be extensively investigated in the past. As a result, the aim of this research was to investigate the different aspects of reliability (test-retest; internal consistency; SEM) of the MABC-2 for age band-2.

Total Test Score

The total test score was examined across three different types of scores with different reliability coefficients. The standard scores were the most reliable in comparison to the component and percentile scores, which was confirmed by the moderate ICC values, and SEM. All scores produced similar ICC values for the TTS, ranging between 0.67 and 0.70, however the SEM is what differentiated the reliability. The SEM of the standard scores was the lowest of the three, as the SEM of the component scores was more than four times as high and the SEM based on the percentile scores was ten times higher. Similarly to the SEM, the SD increased in the same pattern across the three scores. The standard scores had the lowest SD, whereas the percentile scores had the highest. The analysis of the individual performances, through scatterplots, suggested that lack of variability may have contributed to the findings. This is not surprising as the children involved in this study were all typically functioning thus their skill level resulted in similar and relatively proficient performances. In terms of the different types of scores that were analyzed the standard scores produced the highest ICC values and because of this these scores would be most trusted. Overall, the data showed that we can trust the TTS based on the standard scores, but not the component or percentile scores due to the high SEM values.

The results from the present study confirm the results from previous research, indicating that the TTS was a reliable measure of an individual's movement difficulties. A previous study on age band 2 of the MABC-2 confirmed that the component scores were a reliable measure, however the data from our study showed that the standard scores should be used instead of the component scores for a more accurate representation of a child's abilities. Previous literature on other age bands of the assessment tool revealed that the standard scores were reliable, based on high ICC coefficients and low SEM values.

Sub-Components

Three sub-components of the MABC-2 were examined using different types of reliability (ICC, Cronbach's alpha, and SEM). Overall it appears that among the three sub-components, the manual dexterity domain was the most reliable. This was confirmed by relatively moderate ICC values for test re-test, questionable Cronbach's alpha and moderate SEM values, based on the standard scores. The manual dexterity section had the highest internal consistency and test re-test reliability for all of the sub-component. Once again, the standard scores proved to be more reliable than the component or percentile scores.

The balance sub-component had the weakest ICC values across the three types of scores. Similarly, balance had the lowest internal consistency when compared to the manual dexterity and aiming and catching sub-components. The SEM results for the sub-component scores followed a similar pattern to those from the TTS, where the SEM values were the smallest for the standard scores, followed by the component scores, and the percentile scores had the largest SEM. The percentile scores had a SEM that was more than ten times as large as the SEM value for the standard scores, which is the same pattern that the scores had for the manual dexterity

sub-component. The percentile scores also had the lowest reliability coefficient, directly contributing to the high SEM.

The internal consistency for the tasks of age band 2 were low, indicating that the tasks within each sub-component may not be representative of the same domain. However, the SEM for the three sub-components was reported to be acceptable for the standard scores, which indicated that similar scores would be achieved if an individual were to be re-tested. This result was surprising as the reliability coefficients for the standard scores were either questionable or approaching a moderate reliability, however, the SD values for the standard scores were low to moderate and that is what contributed to the acceptable SEM values.

Examination of the individual items of the assessment showed that there were some task specific problems. These problems include the drawing trail-2 task as well as the one board balance task, both which proved to be too difficult for many children and results were not consistent across trials. As well, the balance sub-component is made up of three tasks, two of which (hopping on mats and walking heel to toe on a line) were too simple for the children, and then the aforementioned one board balance task, which was difficult.

These findings were comparable to previous studies that have looked at the test re-test reliability of the sub-components and are therefore robust. All studies reported that the manual dexterity sub-component had the highest reliability, in relation to the other sub-components (Ellinoudis et al., 2011; Smits-Engelsman et al., 2011; Holm et al., 2012; Wuang et al., 2013). Age band 2 of the MABC-2 seemed to show, across all studies, the highest SEM values amongst all age bands of the assessment and this could be due to the lower reliability coefficients for this age band.

Conclusion

The purpose of this study was to examine different facets of reliability (test-retest, internal consistency, and standard error of measurement) of the MABC-2, age band 2 (7 to 10 years old) across different scores (standard, component, and percentile). Overall, low to moderate correlations and generally high SEM across the different types scores indicated that the reliability of this tool is questionable. Teachers, clinicians, and researchers should be hesitant to use this tool for children between these ages until further research is conducted.

The findings of this study were consistent with previous research on age band 2. There has only been one study done, to date, on the reliability of the MABC-2 for age band. The findings from the present research were comparable to the scenario emerging from the previous study examining this age band (Holm et al., 2013). However, the results were not in line with investigations examining the other age bands. This patter of results indicates that the items in age band-2 are less reliable as compared to the other age bands, or tasks within the other age bands. This finding once again emphasizes the importance of examining the different aspects of the tests as some components may be more reliable than others. From a clinical perspective, this is an important finding that should encourage the clinical practitioners to either examine the child with this tool more than once, and/or rely on other sources of information (e.g. Bruininks; parental reports) when trying to infer a child's movement status.

In terms of some shortcomings of this research, the primary limitation of this study was the characteristics and its size. As previously mentioned, most of the children were recruited from sports teams/camps and were therefore physically active and athletic individuals. This created a very homogeneous sample and generated a ceiling effect with some of the tasks, as the participants often found the task to be too easy and would achieve a perfect score. This also

created a lack of variance within the data, as there were not many individuals who scored in the lower percentiles. As well, the study was only conducted on age band 2, which consisted of children between the ages of 7 and 10, with a mean age of 9 years, 0 months and 5 days (SD=1 year, 0 months and 15 days). The results might not be a true representation of the age band as a whole.

Given the scope, the sample was also relatively small. Although the sample size was consistent with previous research conducted by Holm and colleagues (2013) and Smits-Engelsman et al., (2011), a larger, and more diverse, sample size would increase the chances of having a higher reliability. A larger sample size more accurately reflects the population mean and is a better representation of the population as a whole. As well, the sample only consisted of typically functioning individuals, which was done because test re-test reliability examines the consistency/stability of an individual and a main characteristic of atypically functioning individuals in instability.

References

Allen, M., & Yen, W. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Atkinson, G., & Nevill, A. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine, 26*(4), 217-238.

Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowits, J. S., Anderson, G., Atkins, D. C., Carlbring, P., Carroll, K. M., Christensen, A., Eddington, K. M., Enlers, A., Feasles, P. J., Keijsess, G. P., Koch, E., Kuyken, W., Lange, A., Lincoln, T., Stephens, R. S., Taylors, S., Trepka, C., & Watson, J. (2011) Intraclass correlation associated with therapists: Estimates and application in planning physiotherapy research. *Cognitive Behaviour Therapy, 40*, [manuscript].

Barnett, A., & Henderson, S. (1998). *An annotated bibliography of studies using the TOMI/Movement ABC*. London, UK: The Psychological Corporation/Harcourt Brace & Company Publishers.

Barnett, L., Minto, C., Lander, N., & Hardy, L. (in press). Inter-rater reliability assessment using the Test of Gross Motor Development-2. *Journal of Science and Medicine in Sport.*

Brown, T., & Lalor, A. (2009). The movement assessment battery for children – second edition (MABC-2): A review and critique. *Physical and Occupational Therapy in Pediatrics, 29*(1), 86-103.

Bruininks, R. (1978). Bruininks-Oseretsky test of motor proficiency: Examiners manual. Circle

Pines, MN: American Guidance Service.

Bruininks, R., & Bruininks, B. (2005). Bruininks-Oseretsky test of motor proficiency (2nd ed.). Minneapolis, MN: NCS Pearson.

Bruton, A., Conway, J., & Holgate, S. (2000). Reliability: What is it and how is it measured? *Physiotherapy, 86*(2), 94-99.

Burton, A., & Miller, D., (1998). *Movement skill assessment.* Champaign, IL: Human Kinetics.

Carmines, E., & Zeller, R. (1979). *Reliability and Validity Assessment.* Thousand Oaks, CA: Sage Publications.

Caro, T., Roper, R., Young, M., & Dank, G. (1979). Inter-observer reliability. *Behaviour, 69*, 303-315.

Chow, S., Chan, L., Chan, C., & Lau, C. (2002). Reliability of the experimental version of the Movement ABC. *British Journal of Therapy and Rehabilitation, 9*, 404-407.

Chow, S., & Henderson, S. (2003). Interrater and test-retest reliability of the Movement Assessment Battery for Chinese preschool children. *American Journal of Occupational Therapy, 57*(5)*, 574-577.*

Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Croce, R., Horvat, M., & McCarthy, E. (2001). Reliability and concurrent validity of the Movement Assessment Battery for Children. *Perceptual and Motor Skills, 93*, 275-280.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Philadelphia: Harcourt Brace Jovanovich College Publishers.

Darr, C. (2005). A hitch-hikers guide to reliability. *SET: Research Information for Teachers, 3*, 59-60.

Drost, E. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*, 105-123.

Ellinoudis, T., Evaggelinou, C., Kourtessis, T., Konstantinidou, Z., Venetsanour, F., & Kambas, A. (2011). Reliability and validity of age band 1 of the movement assessment battery for children – second edition. *Research in Developmental Disabilities, 32*, 1046-1051.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: L Erlbaum Associates.

Falk, C. F. & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment, 93*(5), 445-453.

Furr, R., & Bacharach, V. (2008). *Psychometrics.* Thousand Oaks, CA: Sage Publications.

Gard, L., & Rosbald, B. (2009). The qualitative motor observations in Movement ABC: Aspects of reliability and validity. *Advances in Physiotherapy, 11*(2), 51-57.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486-489.

Gratton, C., & Jones, I. (2004). *Research Methods for Sports Studies.* New York, NY: Routledge.

Hand, D. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society, Series A, 159*(3), 445-492.

Harvill, L. M. (1991). Standard error of measurement. *Education Measurement: Issues and Practice*, 33-41.

Harwell, M. (2011). *Research in Education – 2nd Edition.* USA: Sage Publications.

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247.

Haywood, K., & Getchell, N. (2009). *Life span motor development – 5th Edition.* Champaign, IL: Human Kinetics.

Henderson, S., & Sugden, D. (1992). *Movement assessment battery for children.* Kent, UK: The Psychological Corporation.

Henderson, S., Sugden, D., & Barnett, A. (2007). *Movement assessment battery for children – 2 second edition* [Movement ABC-2]. London, UK: The Psychological Corporation.

Holm, I., Tveter, A., Aulie, V., & Stuge, B. (2013). High intra- and inter-rater chance variation of the movement assessment battery for children 2, age-band 2. *Research in Developmental Disabilities, 34*(2), 795-800.

Houwen, S., Hartman, E., Jonker, L., & Visscher, C. (2010). Reliability and validity of the TGMD-2 in primary-school-aged children with visual impairments. *Adapted Physical Activity Quarterly, 27*, 143-159.

Hua, J., Gu, G., Meng, W., & Wu, Z. (2013). Age band 1 of the Movement Assessment Battery for Children – Second Edition: Exploring its usefulness in mainland China. *Research in Developmental Disabilities, 34*(2), 801-808.

Hua, J, Wu, Z., Gu, G., & Meng, W. (2012). Assessment on the application of 'Movement Assessment Battery' for Children. *Zhonghua Liu Xing Bing Xue Za Zhi, 33*(10), 1010-1015.

Karpljuk, D., Mesko, M., Videmsek, M., & Tkavc, S. (2009). The relationship between test, measurement and evaluation in human performance. *The International Military Sport Council.*

Kim, H. (2013). Statistical notes for clinical researcher: Evaluation of measurement error 1: Using intaclass correlation coefficient. *Restorative Dentistry and Endodontics, 38*(2), 98-102.

Kirk, J., & Miller, M. (1986). *Reliability and validity in qualitative research.* Beverly Hills, CA: SAGE Publications.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lord, F., & Novick, M (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.

Losardo, A., & Notari-Syverson, A. (2001). *Alternative approaches to assessing young children.* Baltimore, MD: Paul H. Brookes Publishing Co.

Martin, G., & Larson, S.D. (2006). Descriptive statistical and graphical displays. *The American Heart Association, 114*, 76-81.

Mays, N., & Pope, C. (2000). Assessing quality in qualitative research. *British Medical Journal, 320*, 50-52.

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires – 3rd Edition.* New York, NY: Oxford University Press.

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, NJ: Educational Testing Services.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Michell, J. (1990), *An Introduction to the Logic of Psychological Measurement*. Hillsdale, New Jersey: Lawrence Erlbaum.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist*, 36, 3, 211-217.

Miller, M. (2008). Reliability. In N. Sallaind (Ed.), *Encyclopedia of educational psychology* (pp. 847-853). Thousand Oaks, CA: SAGE Publications.

Moore, J., Reeve, T., & Boan, T. (1986). Reliability of the short form of the Bruininks-Oseretsky test of motor proficiency with five year old children. *Perceptual and Motor Skills, 62*, 223-236.

Newell, K. M. (1986). *Constraints on the development of coordination.* Motor Development in Children: Aspects of Coordination. Dordrecht, Germany: Martinus Nijhoff.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory – 3rd Edition*. New York, NY: McGraw-Hill Inc.

Osborne, J. W. (2008). *Best Practices in Quantitative Methods.* Thousand Oaks, CA: SAGE Publications.

Overend, T., Anderson, C., Sawant, A., Perryman, B., & Locking-Cusolito, H. (2010). Relative and absolute reliability of physical function measures in people with end-stage renal disease. *Physiotherapy Canada, 62*, 122-128.

Rodgers, J., & Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*(1), 59-66.

Schoemaker, M., Niemeijer, A., Flapper, B., & Smits-Engelsman, B. (2012). Validity and reliability of the Movement Assessment Battery for Children – 2 Checklist for children with and without motor impairments. *Developmental Medicine and Child Neurology, 54*(4), 368-375.

Schoemaker, M., Smits-Engelsman, B., & Jongmans, M. (2003). Psychometric properties of the Movement Assessment Battery for Children – Checklist as a screening instrument for children with a developmental co-ordination disorder. *British Journal of Educational Psychology, 73*(3), 425-441.

Shultz, K. (2005). *Classical true score theory and reliability.* Sage Publications.

Sim, J., & Wright, C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Journal of the American Physical Therapy Association, 85*, 257-268.

Slacks, M., & Draugalis, J. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy, 58*(22), 2182-2183.

Smits-Engelsman, B., Fiers, M., Henderson, S., & Henderson, L. (2008). Interrater reliability of the movement assessment battery for children. *Physical Therapy, 88*(2), 286-294.

Smits-Engelsman, B., Niemeijer, A., & van Waelvelde, H. (2011). Is the movement assessment battery for children- 2nd edition a reliable instrument to measure motor performance in 3 year old children? *Research in Developmental Disabilities, 32*, 1370-1377.

Steckler, A., & McLeroy, K. (2008). The importance of external validity. *American Journal of Public Health, 98*(1), 9-10.

Stott, D. H., Moyes, F. A., & Headridge, S. E. (1968). Test of motor impairment. Guelph, Ontario, Canada: University of Guelph, Department of Psychology.

Suen, H., & Lei, P.A. (2007). Classical versus generalizability theory of measurement. *Educational Measurement, 4*, 1-13.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55.

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, 6, 35-39.

Trochim, W. M. K. (2005). The theory of measurement. In *Research methods: The concise knowledge base* (pp. 48-74). Cincinnati, OH: Atomic dog.

Venetsanou, F., Kambas, A., Aggeloussis, N., Serbezis, V., & Taxildaris, K. (2007). Use of the Bruininks-Oseretsky Test of Motor Proficiency for identifying children with motor impairment. *Developmental Medicine and Child Neurology, 49*, 846-848.

Venetsanou, F., Kambas, A., Ellinoudis, T., Fatouros, I., Giannakidou, D., & Kourtessis, T. (2011). Can the Movement Assessment Battery for Children-Test be the "gold standard" for the motor assessment of children with Developmental Coordination Disorder? *Research in Developmental Disabilities, 32*, 1-10.

Visser, J., & Jongmans, M. (2004). *Extending the movement assessment battery for children to be suitable for 3-year-olds in the Netherlands.* Unpublished manuscript.

Webb, N., Shavelson, R., & Haertel, E. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistic, 26*, 81-124.

Weir, P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength and Conditioning Research, 19*(1), 231-240.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84*(3), 608-618.

Wiart, L., & Darrah, J. (2001). Review of four tests of gross motor development. *Developmental Medicine and Child Neurology, 43*, 279-285.

Wuang, Y., & Su, C. (2009). Reliability and responsiveness of the Bruininks-Oseretsky Test of Motor Proficiency – second edition in children with intellectual disabilities. *Research in Developmental Disabilities, 30*(5), 847-855.

Appendices

Appendix A

Old vs. New Items of MABC-2 for Age Band 2

Old vs. New Items of MABC-2 for Age Band 2

Table 2b: Brief summary of changes made to AB2 and AB3 – now labelled AB2 and covers ages 7 to 10 years

Task	Movement ABC AB2	Movement ABC AB3	Movement ABC-2 AB2
Manual Dexterity 1	Placing Pegs	Shifting Pegs by Rows	Placing Pegs~
Manual Dexterity 2	Threading Lace	Threading Nuts on Bolt	Threading Lace^
Manual Dexterity 3	Flower Trail	Flower Trail	Drawing Trail 2*
Aiming & Catching 1	Two-Hand Catch	One-Hand Bounce and Catch	Catching with Two Hands
Aiming & Catching 2	Throwing Beanbag into Box	Throwing Beanbag into Box	Throwing Beanbag onto Mat**
Balance 1	Stork Balance	One-Board Balance	One-Board Balance
Balance 2	Heel-to-Toe Walking	Ball Balance	Walking Heel-to-Toe Forwards
Balance 3	Jumping in Squares	Hopping in Squares	Hopping on Mats+

Altered items:

~ New start position/layout

^ Lacing board now longer

* Shape of trail has changed

** Mat with target now used instead of box

+ Mats used for this task

Appendix B

Letter to Thunder Bay Catholic District School Board

Information Letter for Thunder Bay Catholic District School Board

To Whom It May Concern:

My name is Alexandra Boyle and I am a Masters' student at Lakehead University, in the School of Kinesiology. As a requirement of my degree, I have to complete a research project. The study that I am conducting is titled the "Reliability of the Movement Assessment Battery for Children – Second Edition: Age Band 2" and will be completed under the supervision of Dr. Eryk Przysucha, faculty advisor.

The study aims at examining the reliability of the Movement Assessment Battery for Children – Second Edition (MABC-2) for children between the ages of 7 and 10 years old, through several different reliability estimates (test retest; internal consistency and inter and intra-rater reliability). The test has recently been revised in England (Henderson, Sugden, & Barnett, 2007), but there is little information regarding its reliability for children of this particular age group. As this is a new version, little research has been conducted to date, therefore this research is important because it investigates the degree of consistency and stability that the MABC-2 possesses.

When a formal assessment tool is lacking reliability, this indicates that measurement error may be a potential problem. If measurement error is present, this poses a validity issue for the emerging inferences. For example, a child could complete the MABC-2 and receive a score indicating that he/she does not have any movement difficulties. However, if the tool lacks reliability, and the child were to be re-tested, his/her score could fluctuate and indicate that he/she actually does present with mild to moderate movement difficulties. In this example, the measurement error has created a false negative, stating that the child did not have any movement difficulties, when in fact he/she did. Thus, if the MABC-2 demonstrates a high reliability, the emerging inferences from the assessment tool will be valid and we as practitioners can have more confidence in our decisions.

I am writing this letter to ask for your permission to distribute recruitment packages to the schools within the Thunder Bay Catholic District School Board and recruit potential participants for the study. The teachers of the grade 2, 3, 4 and 5 classes will be supplied with the recruitment packages and asked to hand them out to all students. The following week, I would return to the same schools to collect the consent forms from the children that are interested, and then the parents would be contacted. Being involved in the recruitment process for this research study adds little to the job description of the teachers and they are not required to do any extra work. Other than the distribution of the recruitment packages, nothing else will take part at the schools and the teachers will not have to worry about their classes being disrupted.

The recruitment packages contain a recruitment/parent information letter, consent form, and an Exercise and Physical Activity Readiness Assessment for Children and Adolescents

(ExPARA). Prior to the initiation of the study, the children will be asked to complete the ExPARA to ensure that the child is physically able to participate in the study. The child will be asked to attend two sessions, one week apart. Both testing sessions will take place in SB-1028 in the Fieldhouse at Lakehead University, and each session will last approximately 1 hour. The test is composed of three sections: manual dexterity, ball skills and balance. The tasks within each section are relatively simple. For example, they involve cutting out figures, throwing balls at a target or balancing on a board, respectively. They are similar to tasks that a child performs daily, either in school or on the playground. Each testing will be performed individually and conducted by my self. As the primary researcher, I have extensive experience in administering such tests. During the testing, the performance of the child will be videotaped on the first of the two occasions. This is implemented in order to examine one of the aspects of reliability of the assessment tool.

There is minimal psychological or physical harm involved in the child's participation in the study. The child can stop performing the tasks at any time, or withdraw from the study altogether. The child's identity will be anonymous in any of the presentations or publications of the study, as a unique number will be used to replace the child's name. The benefit from participation in this study is the fact that the child will have access to his or her own results and the groups' results. The data will provide information on how well the child is performing the respective tasks against the norms. Also, the board and the schools will be provided with the results and will be mentioned in any presentations and/or publications recognizing their cooperation in the recruitment process. This study has been approved by the Lakehead University Research Ethics Board. If you have any questions related to the ethics of the research and would like to speak to someone outside of the research team please contact Sue Wright at the Research Ethics Board at 807-343-8283 or research@lakeheadu.ca. If you have any additional questions, please do not hesitate to contact myself or my faculty advisor, Dr. Eryk Przysucha.

Thank you for your consideration.

Yours truly,

Alexandra Boyle (aboyle@lakeheadu.ca)

Dr. Eryk Przysucha (Faculty Advisor) (eprzysuc@lakeheadu.ca)

Appendix C

Letter to Lakehead Public School Board

Information Letter for Lakehead Public School Board

To Whom It May Concern:

My name is Alexandra Boyle and I am a Masters' student at Lakehead University, in the School of Kinesiology. As a requirement of my degree, I have to complete a research project. The study that I am conducting is titled the "Reliability of the Movement Assessment Battery for Children – Second Edition: Age Band 2" and will be completed under the supervision of Dr. Eryk Przysucha, faculty advisor.

The study aims at examining the reliability of the Movement Assessment Battery for Children – Second Edition (MABC-2) for children between the ages of 7 and 10 years old, through several different reliability estimates (test retest; internal consistency and inter and intra-rater reliability). The test has recently been revised in England (Henderson, Sugden, & Barnett, 2007), but there is little information regarding its reliability for children of this particular age group. As this is a new version, little research has been conducted to date, therefore this research is important because it investigates the degree of consistency and stability that the MABC-2 possesses.

When a formal assessment tool is lacking reliability, this indicates that measurement error may be a potential problem. If measurement error is present, this poses a validity issue for the emerging inferences. For example, a child could complete the MABC-2 and receive a score indicating that he/she does not have any movement difficulties. However, if the tool lacks reliability, and the child were to be re-tested, his/her score could fluctuate and indicate that he/she actually does present with mild to moderate movement difficulties. In this example, the measurement error has created a false negative, stating that the child did not have any movement difficulties, when in fact he/she did. Thus, if the MABC-2 demonstrates a high reliability, the emerging inferences from the assessment tool will be valid and we as practitioners can have more confidence in our decisions.

I am writing this letter to ask for your permission to distribute recruitment packages to the schools within the Lakehead Public School Board and recruit potential participants for the study. The teachers of the grade 2, 3, 4 and 5 classes will be supplied with the recruitment packages and asked to hand them out to all students. The following week, I would return to the same schools to collect the consent forms from the children that are interested, and then the parents would be contacted. Being involved in the recruitment process for this research study adds little to the job description of the teachers and they are not required to do any extra work. Other than the distribution of the recruitment packages, nothing else will take part at the schools and the teachers will not have to worry about their classes being disrupted.

The recruitment packages contain a recruitment/parent information letter, consent form, and an Exercise and Physical Activity Readiness Assessment for Children and Adolescents

(ExPARA). Prior to the initiation of the study, the children will be asked to complete the ExPARA to ensure that the child is physically able to participate in the study. The child will be asked to attend two sessions, one week apart. Both testing sessions will take place in SB-1028 in the Fieldhouse at Lakehead University, and each session will last approximately 1 hour. The test is composed of three sections: manual dexterity, ball skills and balance. The tasks within each section are relatively simple. For example, they involve cutting out figures, throwing balls at a target or balancing on a board, respectively. They are similar to tasks that a child performs daily, either in school or on the playground. Each testing will be performed individually and conducted by my self. As the primary researcher, I have extensive experience in administering such tests. During the testing, the performance of the child will be videotaped on the first of the two occasions. This is implemented in order to examine one of the aspects of reliability of the assessment tool.

There is minimal psychological or physical harm involved in the child's participation in the study. The child can stop performing the tasks at any time, or withdraw from the study altogether. The child's identity will be anonymous in any of the presentations or publications of the study, as a unique number will be used to replace the child's name. The benefit from participation in this study is the fact that the child will have access to his or her own results and the groups' results. The data will provide information on how well the child is performing the respective tasks against the norms. Also, the board and the schools will be provided with the results and will be mentioned in any presentations and/or publications recognizing their cooperation in the recruitment process. This study has been approved by the Lakehead University Research Ethics Board. If you have any questions related to the ethics of the research and would like to speak to someone outside of the research team please contact Sue Wright at the Research Ethics Board at 807-343-8283 or research@lakeheadu.ca. If you have any additional questions, please do not hesitate to contact myself or my faculty advisor, Dr. Eryk Przysucha.

Thank you for your consideration.

Yours truly,

Alexandra Boyle (aboyle@lakeheadu.ca)

Dr. Eryk Przysucha (Faculty Advisor) (eprzysuc@lakeheadu.ca)

Appendix D

Recruitment/Parent Information Letter

Recruitment/Parent Information Letter

Title of the study: Reliability of the Movement Assessment Battery for Children – Second Edition: Age Band 2

My name is Alexandra Boyle and I am a Masters' student at Lakehead University, in the School of Kinesiology. As a part of my degree, I have to complete a research project. The study that I would like to do is titled the "Reliability of the Movement Assessment Battery for Children – Second Edition: Age Band 2" and will be completed under the supervision of Dr. Eryk Przysucha, faculty advisor.

The study aims at examining the reliability of the Movement Assessment Battery for Children – Second Edition (MABC-2) for children between the ages of 7 and 10 years old, through several different reliability estimates (test retest; internal consistency and inter and intra-rater reliability). These estimates will tell us how consistent this assessment tool is, hence how much faith we can put in its results when assessing children. This test has been around for many years, but recently the authors released a new version (Henderson, Sugden, & Barnett, 2007). Unfortunately, there is still little information regarding its reliability for children of this particular age group.

This issue is very important to researchers and practitioners because when a formal assessment tool is lacking reliability, this indicates that there may be a lot of measurement error. If the error is present then a child who completes the MABC-2 may be judged as having movement problems, while he has none. And, vice-versa, an individual who has problems and needs clinical treatment would score higher than expected thus identifying no problems. As evident, lack of reliability in the scores represents an important issue when assessing children with or without movement problems.

Your child can take part in this study if he/she is between 7 and 10 years of age, and is typically functioning in terms of his/her motor and cognitive status. Atypically functioning children with an official diagnosis for any developmental disabilities (cognitive or motor) will not be considered for this study. This exclusion is due to the fact that a key characteristic of those diagnosed with developmental disabilities is inconsistency, which could affect the results of the study since the consistency of the performance is what is being examined. Prior to the initiation of the study, you will be asked to complete the Exercise and Physical Activity Readiness Assessment for Children and Adolescents (ExPARA) (Appendix E) to ensure that your child is physically able to participate in the study. The ExPARA is a questionnaire that asks general questions about what your child can and cannot do. As a participant in the study, your child will be asked to perform the test twice, one week apart. Both testing sessions will take place in SB-1028 in the Fieldhouse at Lakehead University, and each session will last approximately 1 hour. You will be expected to provide transportation to and from the testing site and a parking pass will be provided for you upon arrival. When you enter the Fieldhouse parking lot at 955 Oliver

Road (across from the Thunder Bay Regional Health Sciences Centre), continue to drive around the building until you reach Lot 1. You will then enter through the blue doors of the building, continue up the stairs and 1028 will be on your left hand side of the hallway. The test is composed of three sections: manual dexterity, ball skills and balance. The tasks within each section are relatively simple. For example, they involve cutting out figures, throwing balls at a target or balancing on a board, respectively. They are similar to tasks that a child performs daily, either in school or on the playground. Each testing will be performed individually and conducted by myself. As the primary researcher, I have extensive experience in administering such tests. During the testing, the performance of your child will be videotaped on the first of the two occasions. This is implemented in order to examine one of the aspects of reliability of the assessment tool. There is an observation room where you can stay for the duration of the testing and observe your child.

There is minimal psychological or physical harm involved in your child's participation in the study. Your child can stop performing the tasks at any time, or withdraw from the study altogether. The child's identity will be anonymous in any of the presentations or publications of the study, as a unique number will be used to replace the child's name. The benefit from participation in this study is the fact that child will have access to his or her own results, and if you or your child wishes, the groups results. The data will provide information on how well your child is performing the respective tasks against the norms.

Following the completion of the study, the information will be stored in a locked filing cabinet or a password-protected computer at Lakehead University for a period of 5 years. If you wish to have access to the results of this study, please include your contact information and a copy of the results will be mailed directly to you. If you consent for your child to participate, please return the attached consent form on the following Monday, as the researcher will be present to collect the forms from those interested. This study has been approved by the Lakehead University Research Ethics Board. If you have any questions related to the ethics of the research and would like to speak to someone outside of the research team please contact Sue Wright at the Research Ethics Board at 807-343-8283 or research@lakeheadu.ca. If you have any additional questions, please do not hesitate to contact myself or my faculty advisor, Dr. Eryk Przysucha

Thank you for your consideration.

Yours truly,

Alexandra Boyle (aboyle@lakeheadu.ca)

Dr. Eryk Przysucha (Faculty Advisor) (eprzysuc@lakeheadu.ca)

Appendix E

Consent Form

Child Participation Consent Form for Parents

Title of the study: Reliability of the Movement Assessment Battery for Children – Second Edition: Age Band 2

I, _____, agree for my child to participate in the research study being conducted by Alexandra Boyle, Master of Science candidate in the school of Kinesiology at Lakehead University.

I have read and understood the information letter for this project. I am aware that there will be two testing sessions, lasting approximately 1 hour each. I agree to complete the ExPARA to ensure that my child is physically able to participate. I understand that the potential risks are minimal and I also recognize the benefits of my child's participation. I am aware that my child's participation is completely voluntary and he/she may withdraw from the study at any given time. I understand that I, or my child, may refuse to answer any questions asked in this research study. I recognize that my child's identity will be anonymous in any of the presentations or publications of the study, as the researcher will use a number to replace my child's name. Dr. Eryk Przysucha will securely store the results of this data in a locked filing cabinet or a password-protected computer for 5 years at Lakehead University. I understand that I may access my child's or the groups' result by contacting the researcher any time after the study is completed.

Participant's Name: _____

Participant's Age & Date of Birth: _____

Parent/Guardian's Name: _____

Signature of the Child: _____

Signature of the Parent/Guardian: _____

Phone Number: _____

Email (optional): _____

Please check this box if you wish to view your child's results

Appendix F

Exercise and Physical Activity Readiness Assessment for Children and Adolescents

Exercise and Physical Activity Readiness Assessment for Children and Adolescents

The purpose of this form is to ensure that we provide every participant with the highest level of care. For most children, physical activity provides an opportunity to have fun and promotes the basis for good health. However there are a small number of children or adolescents who may be at risk when participating in an exercise/ physical activity program. Completion of this questionnaire is mandatory and your child cannot participate in the study entitled "The Reliability of the Movement Assessment Battery for Children – Second Edition: Age Band 2" until it has been submitted. The information contained in this form is confidential and is subject to the regulations of the Privacy Act.

Personal Details

Name:_____ DOB:_____ M/F:_____
Height:_____ Weight:_____
How old was your child as of January 1st? _____
Name of Parent/s or Guardian/s:_____
Phone: home:_____
Has a physician or other medical specialist referred your child?_____
Doctor's Name:_____

Heart-Lung-Other Systems

Does your child have or has had:

A heart condition (please specify)_____
Diabetes (type 1 or 2 - please specify)_____
Cystic Fibrosis:_____ High Blood Pressure:_____ **High Cholesterol:**_____
Breathing problems or shortness of breath (e.g.: asthma)_____
Coughing during or after exercise Other (please specify)_____

Does your child experience or have ever experienced:

Epilepsy or seizures/convulsions: _____
If yes, is it at rest or during exercise?_____
Fainting:_____ Dizzy spells:_____ Heat stroke/heat related illness:_____
Increased bleeding tendency/ haemophilia Other (please specify)_____

None of the above

If your child is taking any medication, please state if there are any side effects experienced as a result of taking this medication: _____

Muscle-Bone System

Has your child ever broken any bones? **Yes** **No**
If so, what bones and when?_____
In the past 6 months, has your child had any muscular pain while exercising? **Yes** **No**
If yes, please explain and indicate where the pain has occurred (e.g. "pain in the back of right heel" or "pain on the inside of the right elbow")

_____
_____H
as a doctor or physiotherapist treated this pain? **Yes** **No**
In the last 6 months, has your child experienced joint pain in the bones? **Yes** **No**
If yes, please explain and indicate where the pain has occurred (e.g.: "front of right leg" or "behind my knee")

_____
_____

Special Conditions

Does your child suffer from any allergies? **Yes** **No**
If yes, please list allergies and any special requirements:_____
Does your child use a "puffer" or "ventilator" for asthma **Yes** **No**
Does your child have any chronic disability or chronic illness? **Yes** **No**
If yes, please indicate condition:

Cerebral Palsy
ADHD
Hypermobility
Intellectual impairment

Are you aware of any medical reason/condition that might prevent your child from participation in an exercise program? **Yes** **No**
If yes, please explain: _____
Is your child participating in any organized sports or extracurricular activities? **Yes** **No**
If yes, what are they? _____
Has your child ever had an operation or injury that required medical intervention? **Yes** **No**
If yes, please explain:_____

Is there anything else that we should know about your child that has not been addressed above?

_____
_____

Informed Consent

I hereby acknowledge that:

- The information provided above regarding my child's health is, to the best of my knowledge, correct.
- I will inform you immediately if there are any changes to the information provided above.
- I give permission for my child to participate in your study.

Parent/Guardian Signature:_____ Date:_____

Approved for participation:_____ Date:_____

Appendix G

Individual Data for Standard Scores

Individual Data for Standard Scores

Participant	Total Test Score – Time 1	Total Test Score – Time 2	Manual Dexterity – Time 1	Manual Dexterity – Time 2	Aiming and Catching – Time 1	Aiming and Catching – Time 2	Balance – Time 1	Balance – Time 2
1	12	13	12	15	8	9	14	14
2	11	12	10	13	9	9	11	12
3	12	11	15	12	8	7	11	14
4	11	12	9	13	10	8	12	12
5	9	9	8	6	8	8	14	14
6	12	11	10	11	12	11	14	9
7	13	12	10	11	14	10	14	14
8	8	8	8	7	5	5	15	15
9	13	8	12	9	11	8	12	9
10	13	12	14	10	9	11	14	14
11	12	10	12	8	8	9	15	15
12	11	12	8	9	14	13	10	14
13	10	9	10	9	8	6	11	11
14	10	14	11	15	8	10	10	12
15	8	9	6	9	11	10	9	8
16	10	10	9	10	8	8	15	15
17	12	15	11	12	12	15	11	12
18	11	11	9	12	10	9	14	11
19	12	11	7	7	14	14	12	14
20	11	11	8	11	12	13	12	9
21	9	11	9	10	10	10	9	12

22	9	11	9	12	9	9	9	11
23	7	10	6	8	10	9	9	14
24	5	10	4	7	8	14	7	10
25	5	5	5	3	9	6	5	8
26	17	16	15	14	15	14	14	14
27	13	13	15	15	8	8	15	15
28	13	16	12	17	11	13	14	14
29	14	14	15	15	10	10	14	14
30	10	11	9	13	12	8	10	10
31	8	10	7	9	12	11	8	10
32	9	8	11	6	7	8	10	10
33	10	13	9	12	8	11	14	14
34	7	9	3	5	14	13	9	9
35	11	12	13	12	7	9	12	12
36	14	15	14	15	10	12	15	15
37	11	15	12	15	10	13	9	12
38	12	16	12	15	12	15	10	11
39	10	10	10	9	9	10	10	11
40	7	8	5	7	8	9	9	10

Appendix H

Individual Data for Component Scores

Individual Data for Component Scores

Participant	Total Test Score – Time 1	Total Test Score – Time 2	Manual Dexterity – Time 1	Manual Dexterity – Time 2	Aiming and Catching – Time 1	Aiming and Catching – Time 2	Balance – Time 1	Balance – Time 2
1	86	92	34	38	16	18	36	36
2	82	88	30	35	18	18	34	35
3	88	83	38	33	16	14	34	36
4	83	86	28	35	20	16	35	35
5	76	73	24	21	16	16	36	36
6	87	82	29	31	22	21	36	30
7	91	87	29	31	26	20	36	36
8	72	70	24	22	11	11	37	37
9	90	72	34	27	21	16	35	29
10	90	87	37	30	17	21	36	36
11	86	79	33	25	16	17	37	37
12	83	86	25	26	26	24	32	36
13	79	74	29	27	16	13	34	34
14	80	94	32	39	16	20	32	35
15	72	74	21	28	21	19	30	27
16	80	81	27	29	16	15	37	37
17	88	97	32	34	23	28	33	35
18	84	85	28	34	20	17	36	34
19	87	85	23	23	26	26	35	36
20	82	83	25	31	22	24	35	28
21	77	85	27	30	20	20	30	35

22	75	84	28	33	17	17	30	34
23	67	78	19	25	20	17	28	36
24	54	80	15	23	15	26	24	31
25	51	51	18	12	18	13	15	26
26	103	99	39	37	28	26	36	36
27	91	92	38	39	16	16	37	37
28	90	101	33	41	21	24	36	36
29	95	95	39	39	20	20	36	36
30	80	82	26	35	23	15	31	32
31	71	80	22	28	23	21	26	31
32	76	68	31	21	14	16	31	31
33	79	91	28	34	15	21	36	36
34	66	73	10	18	26	25	30	30
35	84	86	35	34	14	17	35	35
36	93	98	37	38	19	23	37	37
37	84	98	34	39	20	24	30	35
38	87	99	33	38	23	27	31	34
39	79	80	29	27	18	20	32	33
40	63	71	17	22	16	17	30	32

Appendix I

Individual Data for Percentile Scores

Individual Data for Percentile Scores

Participant	Total Test Score – Time 1	Total Test Score – Time 2	Manual Dexterity – Time 1	Manual Dexterity – Time 2	Aiming and Catching – Time 1	Aiming and Catching – Time 2	Balance – Time 1	Balance – Time 2
1	75	84	75	95	25	37	91	91
2	63	75	50	84	37	37	63	75
3	75	63	95	75	25	16	63	91
4	63	75	37	84	50	25	75	75
5	37	37	25	9	25	25	91	91
6	75	63	50	63	75	63	91	37
7	84	75	50	63	91	50	91	91
8	25	25	25	16	5	5	95	95
9	84	25	75	37	63	25	75	37
10	84	75	91	37	63	25	75	37
11	75	50	75	25	25	37	95	9
12	63	75	25	37	91	84	50	91
13	50	37	50	37	25	9	63	63
14	50	91	63	95	25	50	50	75
15	25	37	9	37	63	50	37	25
16	50	50	37	50	25	25	95	95
17	75	95	63	75	75	95	63	75
18	63	63	37	75	50	57	91	63
19	75	98	75	95	75	95	50	63
20	63	95	75	95	50	84	37	75
21	91	95	91	95	50	75	95	95

22	63	75	84	75	16	37	75	75
23	16	37	1	5	91	84	37	37
24	50	84	37	75	25	63	91	91
25	37	25	63	9	16	25	50	50
26	25	50	16	37	75	63	25	50
27	50	63	37	84	75	25	50	50
28	91	91	95	95	50	50	91	91
29	84	98	75	99	63	84	91	91
30	84	84	95	95	25	25	95	95
31	99	98	95	91	95	91	91	91
32	5	5	5	1	37	9	5	25
33	5	50	2	16	25	91	16	50
34	16	50	9	25	50	37	37	91
35	37	63	37	75	37	37	37	63
36	37	63	37	50	50	50	37	75
37	63	63	25	63	75	84	75	37
38	75	63	16	16	91	91	75	91
39	50	50	50	37	37	50	50	63
40	16	25	5	16	25	37	37	50

Appendix J

Internal Consistency Individual Data for Time One

Internal Consistency Individual Data for Time One

Participant	Manual Dexterity – Task 1	Manual Dexterity – Task 2	Manual Dexterity – Task 3	Aiming and Catching – Task 1	Aiming and Catching – Task 2	Balance – Task 1	Balance – Task 2	Balance – Task 3
1	12	11	11	7	9	13	11	12
2	7	12	11	9	9	11	11	12
3	13	15	11	7	9	11	11	12
4	9	13	6	9	11	13	11	12
5	7	11	6	9	6	13	11	12
6	11	14	4	10	12	13	11	12
7	11	12	6	14	12	13	11	12
8	12	6	6	6	5	14	11	12
9	12	11	11	10	12	12	11	12
10	14	11	12	7	9	13	11	12
11	11	11	11	8	8	14	11	12
12	13	6	6	12	14	9	11	12
13	7	11	11	10	6	11	11	12
14	9	11	12	8	8	8	12	12
15	11	4	6	10	11	7	11	12
16	14	12	11	10	6	14	11	12
17	9	11	12	12	11	10	11	12
18	11	11	6	12	8	13	11	12
19	11	11	11	12	11	8	11	12
20	13	14	6	9	11	11	11	8
21	12	13	12	8	11	14	11	12
22	11	13	11	6	8	12	11	12

23	4	3	3	14	12	7	11	12
24	9	13	6	10	5	13	11	12
25	9	11	11	9	5	8	11	12
26	12	4	6	9	14	8	6	12
27	9	11	6	12	11	8	11	12
28	13	15	11	12	8	13	11	12
29	11	11	11	10	11	13	11	12
30	14	12	12	7	9	13	12	12
31	14	14	11	14	14	13	11	12
32	7	5	6	7	11	4	7	4
33	9	5	1	9	6	8	4	12
34	9	6	4	12	8	5	11	12
35	10	14	4	12	5	7	11	12
36	13	11	3	8	12	6	12	12
37	7	16	2	10	12	11	12	12
38	6	11	6	12	14	12	11	12
39	11	12	6	10	8	9	11	12
40	8	6	3	9	7	7	11	12

Appendix K

Internal Consistency Individual Data for Time Two

Internal Consistency Individual Data for Time Two

Participant	Manual Dexterity – Task 1	Manual Dexterity – Task 2	Manual Dexterity – Task 3	Aiming and Catching – Task 1	Aiming and Catching – Task 2	Balance – Task 1	Balance – Task 2	Balance – Task 3
1	12	14	12	7	11	13	11	12
2	11	12	12	9	9	12	11	12
3	9	13	11	7	7	13	11	12
4	11	13	11	7	9	12	11	12
5	8	7	6	9	7	13	11	12
6	13	13	5	9	13	12	10	8
7	12	13	6	8	12	13	11	12
8	9	7	6	6	5	13	12	12
9	8	13	6	7	9	10	11	8
10	12	12	6	12	9	13	11	12
11	11	10	4	11	6	13	12	12
12	9	11	6	12	12	13	11	12
13	12	12	3	7	6	11	11	12
14	15	13	11	8	12	12	11	12
15	11	12	5	10	9	9	10	8
16	13	11	5	7	8	14	11	12
17	11	11	12	14	14	12	11	12
18	10	12	12	8	9	11	11	12
19	6	11	6	12	14	12	11	12
20	13	13	5	15	9	10	10	8
21	13	11	3	8	12	6	12	12
22	12	15	6	12	5	11	11	12

23	11	10	4	12	5	13	11	12
24	10	9	4	15	11	12	7	12
25	5	6	1	7	6	3	11	12
26	12	14	11	14	12	13	11	12
27	14	13	12	7	9	13	12	12
28	15	15	11	12	12	13	11	12
29	13	15	11	9	11	13	11	12
30	11	12	12	7	8	9	11	12
31	11	12	5	7	14	8	11	12
32	8	7	6	9	7	8	11	12
33	9	14	11	10	11	13	11	12
34	7	7	4	14	11	11	11	8
35	11	12	11	6	11	12	11	12
36	14	12	12	12	11	14	11	12
37	13	15	11	12	12	12	11	12
38	13	14	11	15	12	11	11	12
39	12	10	5	12	8	10	11	12
40	9	8	5	10	7	9	11	12

Appendix L

Supplementary Tables

Supplementary Tables

Intraclass Correlation Coefficients (ICC) Test Re-test Reliability.

	Standard Scores	Component Scores	Percentile Scores
Total Test Score	0.67	0.70	0.68
Manual Dexterity	0.68	0.71	0.64
Aiming & Catching	0.65	0.62	0.63
Balance	0.66	0.55	0.42

Note. The assumptions were not met for the component and percentile scores

Means and Standard Deviations (SD) of Test and Re-Test for Standard, Component, and Percentile Scores.

	Total Test Score		Manual Dexterity		Aiming and Catching		Balance	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MABC-2 – Standard Scores								
• Test	10.55	2.49	9.85	3.10	9.95	2.31	11.55	2.57
• Re-test	11.33	2.53	10.70	3.32	10.13	2.54	12.10	2.18
MABC-2 – Component Scores								
• Test	80.78	10.51	28.5	6.86	19.28	3.96	32.93	4.36
• Re-test	83.73	10.34	30.30	6.79	19.48	4.30	33.95	2.96
MABC-2 – Percentile Scores								
• Test	56.20	25.64	48.93	29.93	49.28	25.34	65.03	25.64
• Re-test	63.00	24.47	56.08	31.76	50.05	27.17	67.88	24.18

Note. The assumptions were not met for the component and percentile scores, therefore the standard scores are the focus of the study.

Cronbach's Alpha for Time One and Time Two of the Subcomponent Scores.

Sub-component	Time One	Time Two
Manual Dexterity	0.61	0.75
Aiming and Catching	0.49	0.42
Balance	0.53	0.26

Note. The results for time one were the primary focus of this study and the Cronbach's alpha values for time two were included for comparison and discussion purposes.

Cronbach's Alpha with Items Deleted for Time Two.

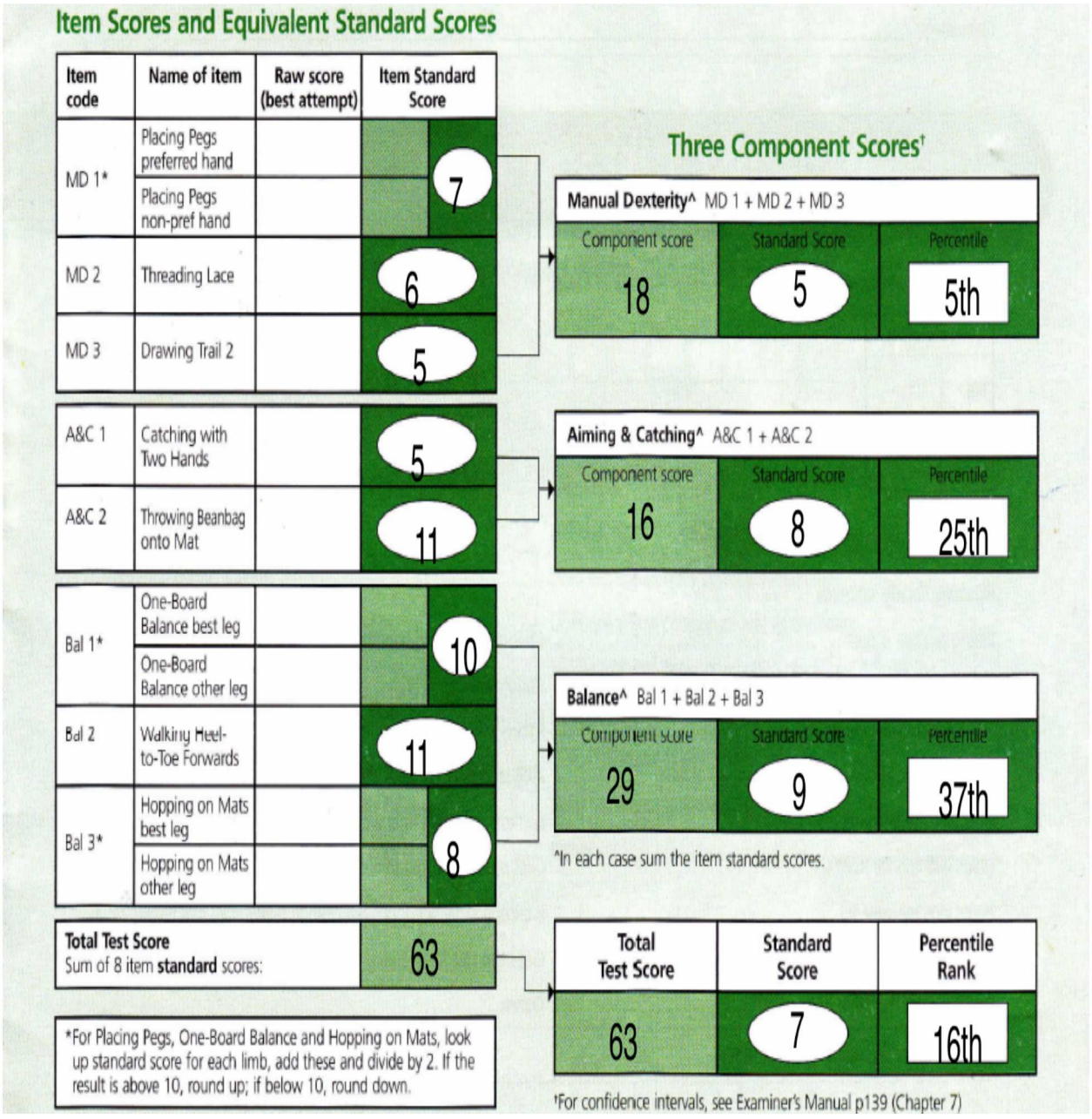
Item	Cronbach's Alpha if Deleted
Manual Dexterity – Cronbach's Alpha = 0.75	
Manual Dexterity 1	0.71
Manual Dexterity 2	0.52
Manual Dexterity 3	0.79
Aiming and Catching – Cronbach's Alpha = 0.42	
Aiming and Catching 1	-
Aiming and Catching 2	-
Balance – Cronbach's Alpha = 0.26	
Balance 1	0.37
Balance 2	0.25
Balance 3	0.002

Note. There are no values for aiming and catching with items deleted, as an item cannot be deleted since there are only two items in this domain.

Appendix M

MABC-2 Scoring Form

MABC-2 Scoring Form



Item Scores and Equivalent Standard Scores

Item code	Name of item	Raw score (best attempt)	Item Standard Score
MD 1*	Placing Pegs preferred hand		7
	Placing Pegs non-pref hand		
MD 2	Threading Lace		6
MD 3	Drawing Trail 2		5
A&C 1	Catching with Two Hands		5
A&C 2	Throwing Beanbag onto Mat		11
Bal 1*	One-Board Balance best leg		10
	One-Board Balance other leg		
Bal 2	Walking Heel-to-Toe Forwards		11
Bal 3*	Hopping on Mats best leg		8
	Hopping on Mats other leg		
Total Test Score Sum of 8 item **standard** scores:			63

*For Placing Pegs, One-Board Balance and Hopping on Mats, look up standard score for each limb, add these and divide by 2. If the result is above 10, round up; if below 10, round down.

Three Component Scores†

Manual Dexterity^ MD 1 + MD 2 + MD 3

Component score	Standard Score	Percentile
18	5	5th

Aiming & Catching^ A&C 1 + A&C 2

Component score	Standard Score	Percentile
16	8	25th

Balance^ Bal 1 + Bal 2 + Bal 3

Component score	Standard Score	Percentile
29	9	37th

^In each case sum the item standard scores.

Total Test Score	Standard Score	Percentile Rank
63	7	16th

†For confidence intervals, see Examiner's Manual p139 (Chapter 7)