The validity of swimming rubrics for children with and without a physical disability

Tae-Sang Jin
Department of Kinesiology and Physical Education
McGill University
Montreal, Canada

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Masters of Arts

Your file Votre référence
ISBN: 978-0-494-32529-2
Our file Notre référence
ISBN: 978-0-494-32529-2

Abstract

The purpose of this study was to investigate the psychometric properties of swimming rubrics. The 10-level rubrics were designed to assess the front crawl. Participants were children, aged 8 to 13 years, with and without a physical disability (n=19) from a "reverse integration" school in Montreal. Participants swam 20 meters with each deciding if a floatation device was necessary. They evaluated themselves as well as peers using the rubric format. The physical education teacher and two teaching assistants participated as teacher assessors. Teacher, peer, and self assessments produced similar scores. In peer assessment, students with disability produced lower scores than students without disability. Boys did not differ from girls. In self assessment, students with and without a disability showed similar competence in comparison to teachers. Also, boys and girls produced similar competence in comparison to their teacher as well. Finally, video assessment was significantly correlated with assessment done immediately after performance.

Abstrait

Le but de cette étude était d'enquêter sur les propriétés psychométriques des rubriques de natation. Des rubriques à 10 niveaux ont été crées pour évaluer le crawl. Les participants étaient des enfants âgés entre 8 et 13 ans, avec ou sans handicap physique (n=19) et qui venaient d'une école à « intégration inversée » de Montréal. Les participants nageaient 20 mètres avec chacun décidant si une bouée de flottaison était nécessaire. Ils étaient évalués par eux-mêmes ainsi que par leurs pairs avec l'aide des rubriques. Le professeur d'éducation physique et deux assistants d'enseignement participaient en tant qu'évaluateurs professoraux. Tous, c'est-à-dire le professeur, les assistants, les pairs et les propres évaluations des participants ont produit des scores similaires. Dans les évaluations de pairs, les étudiants avec des handicaps ont produit des scores plus bas que les étudiants sans handicap. Il n'y a pas eu de différence entre garçons et filles. Dans les propres évaluations, les étudiants avec et sans handicaps ont montré des compétences similaires en comparaison avec les professeurs. Ainsi, les garçons et les filles ont produit des compétences similaires en comparaison avec le professeur aussi. Finalement, les évaluations par vidéo étaient significativement en corrélation avec l'évaluation faite immédiatement après la performance.

Acknowledgment

My sincere thanks go to my academic supervisor, Dr. Greg Reid, for his patient support and guidance. He was not only my academic supervisor but also a kind and prominent English writing teacher. Consequently, his endless and positive reinforcement made this all possible.

Next, I would like to acknowledge the contribution of Helena Seymour who assisted in conducting this research by creating swimming rubrics, making videotapes, and participating in research activities.

I also would like to thank all the staff at the Mackay Center School, especially Bob Simpson, physical education teacher, for his enormous support by sharing the class times and contributing with ideas of developing swimming rubrics.

Another appreciation goes to my beautiful wife, Hyun-Hui Chung. She has endured a long tunnel of tough life with our newborn baby in this foreign country. Her encouragement revived my distressed heart to carry out this work.

And finally, to the One who is my Lord and my strength – Jesus Christ. Throughout tough times studying in Canada in a second language, You were the only One whom I relied on, and I will hold onto throughout my whole life.

Table of Contents

Chapter 1

INTRODUCTION

Fundamentally, education begins at birth and continues throughout the lifespan. Philosophies have changed and evolved in accordance with social demands of different generations and new education practices. According to the United Nations, the primary goal of education is to develop a child to his or her full potential (http://www.unac.org/en/link_learn/monitoring/Childrights_education.asp). Included in this goal is respect for human rights, sense of identity and affiliation, and interaction with others and the environment. Wehmeyer, Sands, Doll, and Palmer (1997) elaborated the goals of education, (1) to produce responsible and self-sufficient individuals who (2) possess self-esteem, initiative, skills, and wisdom to who (3) continue individual growth and pursue knowledge. Specifically, Wehmeyer and his colleagues underlined that students with disabilities should become more self-sufficient and self-determined to achieve these educational goals (Wehmeyer et al., 1997).

Despite the goals of education, a number of professionals have expressed dissatisfaction with classroom, and educational environments (Janesick, 2001; Montgomery, 2001; Wiggins, 1998). For example, teacher-centered instructional methods cause students to be passive and less engaged in their own learning, regardless of the child's characteristics and capabilities. In accordance with this approach, teachers often assess children with standardized tests, which generally identify correct responses only, and do not necessarily determine what a child really knows and is capable of performing in the real world (Janesick, 2001). In addition, many education personnel use inappropriate assessment tools for the exclusive purpose of gathering data in order to benefit from the advantages of statistical accuracy and economy (Wiggins, 1989).

Furthermore, schools have been criticized for focusing solely on strategies to enhance test scores, and for mechanically training their students to learn strictly by memory in order to achieve better grades in high-stake tests.

Today's learners need problem-solving skills and ways of thinking critically, and to select and utilize the information they obtain. Furthermore, schools should strive to create opportunities for students to display what they have learned to verify their progress (Janesick, 2001). In addition, students should be encouraged to demonstrate self-determination. To summarize, students need to be aware of their personal needs, set self-determined goals, and persistently pursue their goals by adjusting purposeful performance and problem solving skills (Martin & Marshall, 1995). Thus, regardless of ability or disability, they should be psychologically empowered to have control over their learning, act autonomously, and regulate their schooling to accommodate their personal preferences and interests (Wehmeyer et al., 1997).

Assessment is critical to education and can be defined as a process of collecting and analyzing data relevant to the characteristics of people, objects, or processes (Chatterji, 2003; Salvia & Ysseldyke, 2004). Educational assessment should be on-going and continuous, providing feedback to learners. Moreover, the learning of students should be improved through assessment, by motivating students to achieve higher levels (Doolittle & Fay, 2002; Hensley, 1997). Wiggins (1989) describes authentic assessment as tasks that resemble real-life settings rather than ones that are artificial, contrived, and typically found in standardized testing. While traditional assessment audits students' discrete knowledge with isolated skills or drills, authentic assessment is deliberately designed to educate and enhance student performance (Wiggins, 1998). It is designed not only to assess but also to improve student's

performance, which includes higher levels of cognitive thinking, such as problem solving and critical thinking (Montgomery, 2002; Wiggins, 1998). Authentic assessment rejects the assumption that scoring high on multiple-choice tests determines that the student is knowledgeable. Rather, it directly assesses students by encouraging them to show progress upon ecologically valid responses. Accordingly, the Education Ministry of Quebec provides guidelines for new competencies for teachers in which assessment takes place on a daily and authentic basis (Government of Quebec, 2001).

There are a number of significant reasons for introducing authentic assessment in schools. First, it is important for students to apply knowledge in practical, real world settings (Wiggins, 1993). To emphasize the differences between 'knowing' and 'applying', Neisser divided intelligence into academic intelligence, and practical intelligence (cited by Sternberg, Wagner, Williams, & Horvath, 1995). This can be described as the difference between how to watch and how to perform. For example, watching a game of ice-hockey involves knowing the rules, the players on each team, and the record and history of their previous encounters. However, to play the game well, it is crucial to have experienced the real game. A person with physical disabilities may not be able to play hockey yet can watch a hockey game; similarly, a person may be able to play the game well yet not be able to follow a game as a spectator.

Second, authentic assessment evaluates, regardless of one's race, class, and gender (Janesick, 2001; Wiggins, 1989). Often, standardized tests do not recognize the society in which a variety of races and cultures co-exist. Wiggins (1989) supports 'ecological validity,' which challenges whether content and procedures are well designed for participants' social and cultural back ground. For example, if an elementary student in Canada was questioned in geography class about the largest South Korean natural

resource, one could argue that ecological validity was not met.

Third, authentic assessment has the merit of continuously assessing students (Block, 2000; Lieberman & Houston-Wilson, 2002; Popham, 2005; Wiggins, 1993). Authentic assessment is not created to evaluate a final product. Traditional evaluation methods such as norm-referenced tests collect scores in order to average them for ranking students, usually at the beginning or end of instruction. This method may be appropriate for placement, but it is not an appropriate method to ameliorate interactive teaching-learning activities, because it cannot identify student weaknesses or strengths in a specific unit of the school curriculum. Therefore, authentic assessment can improve the teaching-learning context.

Fourth, authentic assessment can have positive effects on learning. Authentic assessment does not encourage dualism in learning where students are pressed to memorize simple information, but encourages students to apply knowledge in varying degrees (Hensley, 1997). Furthermore, it is often argued that traditional assessment via techniques such as multiple-choice exams fail to measure learning retained after the exam, while authentic assessment can track on-going learning of students (Kirk, 1997). For instance, creating a portfolio, whereby student skill development is exhibited is an appropriate method for observing student progress and status.

Educational assessment is an integral part of the instructional process. Instruction and curriculum should correspond with the objectives of the school's educational system. If a school's educational objective is to pass standardized tests, then the majority of students will be instructed to memorize information in preparation for the test, and then search for the correct response amongst the list of multiple choice answers when taking the tests. As noted, educational assessment should be on-going

and provide feedback to learners. Wiggins (1989) has argued that authentic assessment not only verifies the students' achievement, but also shows the actual challenges and progress of the student.

Like art (portfolios) and music (performances), authentic assessment has been widely adopted for evaluation purposes in the context of physical education. Students in physical education are often asked to perform certain tasks to demonstrate their ability. However, historically physical education instruction has focused on practicing drills and isolated skills. Skill tests are indicators of a student's skill level but may not predict the student's ability to be successful in game play (Lund, 1997). Thus, the level of assessment must be more holistic for the task to be considered significant and authentic.

The National Association for Sport and Physical Education (NASPE) published a landmark document, *Moving into the Future: National Standard for Physical Education* (1995), which provides resources, including current, appropriate, and realistic assessment tools. An abundance of books and documents offer scoring rubrics as an effective type of authentic assessment in physical education (Block, Lieberman, & Connor-Kuntz, 1998; Doolittle & Fay, 2002; Gibbons & Robinson, 2005; Hensley, 1997; Lieberman & Houston-Wilson, 2002). Rubric systems can be found in many sports activities, such as in martial arts or in swimming. For instance, karate and Tae Kwon Do both employ the use of different color belts as a system for classifying different levels of performance. Similarly, Aqua Quest, a swimming guideline set by the Canadian Red Cross Society, divides the levels into twelve, in teaching swimming.

Moreover, many of rubrics can be found on the Internet sites such as http://www.rubistar.org/, http://www.teach-nology.com/web_tools/rubrics/, http://www.cotf.edu/ and http://school.discovery.com/schrockguide/assess.html.

A rubric is a carefully described scoring system with specific standards and criteria for judging student performance (Smith & Cestaro, 1998), and directly guiding focus on key elements as students work toward mastery (Lund, 1997). This new assessment tool has gained popularity in education and physical education as a way for students to understand what teacher expects and, in turn, for teachers to assess students' products, progress, and the process of learning (Gibbons & Robinson, 2005; Hall & Salmon, 2003; Lieberman & Houston-Wilson, 2002; Montgomery, 2000). Significant advantages for using rubrics for assessment have been noted. First, rubrics allow teachers to teach and assess at the same time, since the rubrics are directly related to instruction (Goodrich, 1997; Hall & Salmon, 2003; Montgomery, 2002). Second, rubrics can accommodate heterogeneous classes such as those that include students with disabilities since they can be expanded and specialized (Block et al., 1998; Goodrich, 1997).

Third, and most significantly, rubrics can help learners to be more self-determined in their learning. Rubrics allow a simultaneous process of teacher, peer and self assessment, providing students with the opportunity to assess themselves and their peers in an interactive manner and in natural environment. Students can even participate in designing and applying rubrics (Woods & Anderson, 2002). The role of the educator changes from that of a teacher, to a supervisor, to an ally by using rubrics. Therefore, rubrics expect students to be held accountable for their own development, through motivation, and being challenged in interesting ways (Lieberman & Houston-Wilson, 2002). In adapted physical education, equal-status relationships are keys to integration or inclusion (Sherrill, Heikinaro-Johansson, & Slininger, 1994). It is true that many students with disabilities accompany their able-bodied peers, but with limited

contact (Sherrill et al., 1994). Peer tutoring could be a strategy to give them opportunities to connect with peers in integrated physical education classes (Houston-Wilson, Dunn, Mars, & McCubbin, 1997; Lieberman, Dunn, Mars, & McCubbin, 2000; Webster, 1987). However, students with a disability often just receive instructions from students without a disability (Lieberman & Houston-Wilson, 2002). This is not a reciprocal equal-status relationship. Rubrics might be a solution to fill the gap, if students, regardless of disability or gender, can accurately assess peers and themselves like teachers.

A study for teacher competency using scoring rubrics was implemented and reported that physical education teachers were as reliable observers (86.89%) as a committee composed of professionals and university faculty members. (Williams & Rink, 2003). However, no empirical data have been reported on the reliability or validity of rubrics in terms of different assessors (teacher, peer, and self) in a physical education and activity setting. Therefore, the purpose of this study was to investigate the psychometric properties of rubrics in an integrated swimming class. Swimming was selected because it is one of the most popular physical activities for all people, including those with disabilities. The buoyancy of water makes swimming an excellent exercise that encourages individuals with physical disabilities to move their body and to develop physical health (Katz & Bruning, 1981).

Statement of problem

The purposes of this study were to compare the use of rubrics (1) by teachers, peers, and self (2) by students with and without a disability, and (3) male and female students.

Hypotheses

1) Teachers, peers, and self will produce similar assessment results.
2) Disability does not affect assessment results.
3) Gender differences do not affect assessment results.

Delimitations

1) Children were recruited from only one school in Montreal, Canada.
2) Children's ages were from 8 to 13 years.
3) Swimming was restricted to the front crawl.

Limitations

This study has a number of limitations:

1) Only two weeks of rubric practice was given to the participants. It is possible that the results reflect, in part, the relative novelty of the rubrics. This might be positive in as much as a "new" approach attracts the attention of the students. It might also be negative in as much as participants must learn how to use the rubrics.
2) Participants had to reach 80% of criterion in using the rubrics in order to qualify as assessors. Only children with a disability were disqualified because they failed to reach 80%. It would seem that some of them need more practice in using the rubrics.
3) Video assessment may not be completely authentic. The front crawl of the participants was captured on video for later assessment by teachers, peers, and

self. This form of video assessment may not be completely authentic since it is an assessment some time after performance.

Chapter 2

LITERATURE REVIEW

The purpose of this study was to investigate the psychometric properties of rubrics in an integrated swimming class. This chapter is divided into 1) Assessment 2) Philosophical background of authentic assessment, and 3) Authentic assessment and rubrics.

Assessment

Definition of assessment

Assessment is the process of collecting and analyzing data which involve knowledge, skills, attitudes, and beliefs in measurable terms (Chatterji, 2003; Salvia & Ysseldyke, 2004). Assessment is often used in an educational context, but it applies to other fields as well, such as health and finance. Several terms, such as measurement, evaluation, and test are similar, or closely related to assessment. Chatterji (2003) argued that "measurement" and "assessment" are often used synonymously. However, he added that the term measurement is more closely related to traditional, standardized achievement tests, while assessment is a broader educational construct, from historical information, to multiple choice tests, to portfolios. All assessment/measurement procedures must be based on professionally established standards of quality for achieving information (Chatterji, 2003).

Chatterji and others (e.g. Salvia and Ysseldyke, (2004)), however, made the clear distinction that "evaluation" is a process after assessment is completed. Evaluation involves a judgement or interpretation of collected data from assessment for making decisions. For example, given a result of 15 seconds for 100M dash, the result may be

evaluated as "good," "normal," or "slow." Therefore, evaluation often accompanies contexts and degree of subjective judgment. In short, evaluation involves decision-making and interpretation of the information obtained from one or more assessments (Chatterji, 2003).

Testing consists of a particular set of questions to an individual or group to obtain a score. Testing is not synonymous with assessment; rather it is one part of the assessment process, which serves to understand students' performance, skills, or knowledge (Chatterji, 2003).

Popham (2005) defined educational assessment as a formal attempt to determine student status with respect to educational variables of interest. It is true that we often judge an individual informally. For example, we may conclude that a child who acts very harshly with a parent is spoiled, or a teacher may conclude that a student is clumsy based on observing one aspect of physical activity. Educational assessment should yield "formal" information to obtain an estimate of a students' status. Tests can help teachers to consider their students formally. However, Kubiszyn and Borich (2003) indicated three concerns when only using test results for judging students. First, tests can be unintentionally misused and intentionally abused. Second, tests can be poorly designed. Finally, ill-trained or inexperienced test administers may perform the test poorly. Moreover, focusing only on the test itself cannot detect what students really know, because student performance is influenced by the (1) task itself, (2) performer's history and characteristics to the task, and (3) context where the test is conducted (Salvia & Ysseldyke, 2004). Thus, educators have been urged to employ a wider variety of measuring devices that cover various educational interests. Therefore, test tools must be valid and reliable in accordance with the context of educational environments and

purposes of testing. Test results should be considered part of the assessment process, and assessment is a broad concept that embraces diverse kinds of tests (Kubiszyn & Borich, 2003; Popham, 2005).

Purposes of assessment

Purposes of assessment vary with regard to who uses the assessment information and what educational decision they make (Bouffard, 2003; Chatterji, 2003; Salvia & Ysseldyke, 2004). First, classroom teachers assess students in classrooms. Effective use of classroom assessment by teachers can facilitate various aspects of teaching and the learning process. Classroom teachers may establish individual student goals, which are composed with long- and short-term outcomes. To achieve those educational goals, teachers need to understand the heterogeneity of students. When curriculum and lessons are planned, both instruction and assessment should be developed so that students clearly understand the desired learning outcomes (Chatterji, 2003).

Second, program developers, managers, administrators, or policymakers are also interested in program-level assessments. Program-level assessments are large-scale assessments that chart educational accountability and high stakes decisions in educational programs. In education, accountability refers to the responsibility of accomplishing goals of an educational institution. Schoolwide accountability usually involves high stakes testing results, where individual staff members are often dependent on the test results. Stakeholders demand better results of educational institutions or funding may be discontinued (Chatterji, 2003).

Third, counselling psychologists, special educators, therapists, school nurses, and social workers are responsible for screening and diagnosis. These personnel use a

variety of clinical and psycho-educational assessments to screen individuals for further assessment, to diagnose particular conditions, and to determine eligibility for therapeutic or special intervention services. It is typically associated with decisions for placement, so the assessment tools must have adequate psychometric credibility and the assessment should be conducted by well trained professionals (Chatterji, 2003).

Fourth, educational decisions made using assessment scores has to do with admission, licensure, promotion, and/or recognition of individuals in an institution, program, or profession. For example, in the area of collegiate admissions, Scholastic Assessment Tests I and II (SAT I, SAT II), the Graduate Record Examination (GRE), and the Law Schools Admission Test (LSAT) are typically used by admission boards in undergraduate, graduate, and professional schools. Test of English as a Foreign Language (TOEFL) is mandated to all international students who intend to be enrolled in an English based school. Although there are many other non-test ways to select individual candidates (e.g., interviews, performance in particular settings, or writing samples), tests are cost-effective and can classify many people more easily and efficiently. These kinds of tests must have technical defensibility in at least three areas (Chatterji, 2003). First, 'predictive validity' should be achieved. For example, users of the TOEFL would be interested in whether the TOEFL scores of an international student actually correlate with his/her later performance in an English school. Second, the assessment data should be free from possible selection biased towards a particular gender, ethnic, or minority groups. Third, the standards or cut-scores used in making selections should be reasonable. Assessment test users should not set scores arbitrarily, but after consideration of statistical and other evidence that the required score is reasonable for the population or subpopulation of interest.

Types of assessment

Summative and formative assessments. Summative assessments are comprehensive in nature, provide accountability, and are generally used to check the level of learning at the end of a course or program. In an educational setting, summative assessments are typically used to assign students a final grade, which is a means of accountability. Program goals and objectives often reflect the cumulative nature of the learning that takes place. Thus, a program would conduct summative assessment at the end to ensure students have met the program goals and objectives. Given too much focus on summative assessment in classrooms, it may solicit teachers to pay attention on preparation for tests, may promote cheating, and not provide information for correcting errors (Wiggins, 1989, 1998).

Formative assessment, however, is used to aid learning and is generally carried out throughout a course or program. Classroom assessment is one of the most common formative assessments. The purpose of this assessment is to improve quality of student learning, mostly by providing feedback on work produced, which allows students to correct conceptual errors. It would not necessarily be used for grading purpose. Thus, this technique prevents motivation of cheating and promotes active reflection on the effectiveness of instruction. While summative assessment is referred to as "assessment of learning," the reference "assessment for learning" is for formative assessment (http://en.wikipedia.org/wiki/Assessment). Black and Wiliam (1998) encourage teachers to use questioning and classroom discussion in order to increase their students' knowledge and improve understanding. They added that tests can be used formatively if teachers analyze where students are in their learning and provide specific, focused feedback (Black & Wiliam, 1998). Diagnostic assessment is a common form of

formative assessment. Diagnostic assessment measures a student's current knowledge and skills for the purpose of identifying a suitable program of learning.

Objective and subjective assessments. Summative and formative assessment can be objective or subjective. Objective assessments are forms of questioning which have a single correct answer. True/false test, multiple choice or multiple-response test, and matching questions are included in objective assessment categories. Today, because of an advantage of quick and easy data collection and analysis, objective assessments are becoming more popular, especially with online assessment, because this form of questioning is well-suited to computerisation. In the education context, however, the score obtained from objective assessments might not always indicate what students really know, since test items are often too simplified and isolated (Wiggins, 1998). Moreover, objective questions can be answered through a guessing strategy.

Subjective assessments are forms of questioning which may have more than one current answer or more than one way of expressing the correct answer. Subjective questions included extended-response questions, essays, and oral tests. This type of assessment can check beyond superficial knowledge. But it is crucial that assessment items should be equipped with clear and appropriate criteria (Gratz, 2000).

Norm-referenced and criterion-referenced tests. Norm referenced assessment tells us where a student stands compared to other students. Data from norm-referenced assessment determines a student's "place" or "rank." The best know example of norm-referenced assessment is the IQ test. Many entrances tests to prestigious schools or universities are also this type of assessment that decide a fixed number of students to enrol. It shows the comparability of selected students rather than an explicit level of their academic ability. Therefore, standards may vary from year to year, depending on

the quality of the cohort.

The other type of assessment, criterion referenced assessment, occurs when candidates are measured against defined criteria. This type of assessment tells us about a student's level of proficiency inm, or mastery of some skill or set of skills (Kubiszyn & Borich, 2003). For example, the driving test is the best known of criterion referenced assessment when student drivers are measured against a range of explicit criteria. Such information helps instructors decide whether a student needs more work on some skill or set of skills. Standards of criterion referenced assessment do not vary from year to year unless the criteria change.

Validity

When collecting research data, validity is one of fundamental criteria for judging the quality of measures. Validity refers to the soundness of the interpretation of a test, the most important consideration in measurement. An historical view of validity is one which indicates the degree to which the test or instrument measures what it is intended to measure (Kubiszyn & Borich, 2003; Thomas & Nelson, 2001). For instance, it would not be valid to assess driving skills through a written test alone; the most valid way of assessing driving skills would be through a combination of practical assessment and written test. It cannot be said that an assessment tool is valid and another tool is not valid. Many interpretations through different contexts could be drawn from the result of an assessment tool, which make validity a matter of degree, not 'black or white' (Bouffard, 2003; Salvia & Ysseldyke, 2004). Therefore, some scholars argued that validity refers to the "plausibility of inferences" (Bouffard, 1993; Yun & Ulrich, 2002). Measurement validation is not a one-time responsibility of the test developer, but it is on-

going process for the dynamic interaction amongst test participants, instrument, context, and purpose of measurement (Yun & Ulrich, 2002). For example, a given test might be valid as a diagnostic tool but not valid for measuring student learning. That is, the inferences from test scores with regard to who has, or does not have, a given condition are plausible if the test is valid.

To summarize, all validity questions should ask whether the assessing process leads to appropriate inferences about a specific person in a specific situation for a specific purpose. There are four major types of validity issues: logical validity, content validity, criterion-related validity, and construct validity.

Logical validity. Logical validity is usually referred to as face validity (Thomas & Nelson, 2001). Logical validity is claimed when the measure obviously involves the performance being measured. In other words, it means that the test is valid by definition. For example, a speed-of-movement test, in which the person is timed in running a specified distance, is considered to have logical validity. Even though logical validity is used in a research study, measurement experts prefer to have more objective evidence of the validity measurement (Thomas & Nelson, 2001). Furthermore, logical validity should be avoided with indirect measures such as attitude, opinion, and/or personality traits, because context and environment may influence the response of participants, and it becomes difficult to differentiate the influence of the context and the true personal traits (Yun & Ulrich, 2002).

Content validity. Content validity evidence is a minimum requirement for a useful test, and often is the only validity evidence we can feasibly gather to support interpretations made from classroom assessment results (Chatterji, 2003; Kubiszyn & Borich, 2003). As with logical validity, no statistical evidence can be supplied for

content validity (Thomas & Nelson, 2001). An assessment tool is developed for a specific purpose. Test items of the assessment tool must represent the domain or universe of the specific purpose. For instance, if it is supposed to assess third-grade arithmetic ability, it should measure third-grade arithmetic skills, not fifth-grade arithmetic skills and not third-grade reading ability. In the physical education context, measuring general motor ability does not predict if one is a very good baseball player. More information is needed about throwing and catching, running, team work, knowing rules of baseball etc that combined make a good baseball player.

In order to provide content-related evidence, frequently test developers ask panels of experts for judgments about the appropriateness of test contents (Bouffard, 2003; Salvia & Ysseldyke, 2004; Yun & Ulrich, 2002). However, it can be difficult to ascertain who is a true content expert who understands the entire domain of the test tools of study, because we do not determine how many years of experience and what level of productivity and/or recognition by peers is required to be a expert of a certain domain. It is a serious concern if the test instrument is developed to make important decisions that may result in serious consequences. Yun and Ulrich (2002) suggested two different strategies that researchers can employ when selecting a panel of experts. First, inviting two different groups of experts (e.g., academic background and practical background) brings two different perspectives on the content evaluation. However, this approach could be limited because of lack of understanding of the entire domain of interest. The second approach is to identify essential content areas and invite content experts who understand these content areas. The selected experts evaluate each developed assessment item according to specific criteria concerning relevance and accuracy, and they summarize information and/or select the final items. Besides the agreement of a

panel of experts, content-related evidence should include an operational definition of the content domain, qualifications of the experts, specific directions given to experts for evaluating the instrument, and specific criteria for selecting the final items (Yun & Ulrich, 2002).

Content-related information has limitations. The major limitation is that content validation only provides evidence of validity of test items, not inferences made from the results of measurements, in which there is dynamic interaction among participants, test instrument, and context (Messick, 1989; Yun & Ulrich, 2002).

Criterion-related validity. Criterion-related validity is the matter of how scores from a test are correlated with an external criterion (Chatterji, 2003; Kubiszyn & Borich, 2003; Salvia & Ysseldyke, 2004). There are two major sources of criterion-related evidence: Concurrent and Predictive validity. Both sources are primarily determined in the temporal relationship between the test scores and the outcome criteria. Concurrent validity indicates the relationship between the test scores and outcome criteria that are analyzed at the same time. In contrast, predictive validity assesses the degree to which test scores are related to outcomes at some point in the future. Although there are differences between concurrent and predictive related validity evidence, both types of criterion-related validity conceptually share the same logic, which is how accurately criterion outcomes are predicted from the test scores (Yun & Ulrich, 2002). Both concurrent and predictive criterion-related validity evidence yield numerical indices of validity, unlike content validity.

Concurrent criterion-related validity concerns whether results of a test can be equivalent to the results on another test (external criterion) at the same time. A new test may have some practical advantages; cheaper, shorter using new technology, or can be

administered to groups (Kubiszyn & Borich, 2003). However, the new test will be not useful or trustworthy if concurrent validity is not established. For instance, the Stanford-Binet and the Wechsler Intelligence Scale for Children-III (WISC-III) are well known, widely accepted IQ tests. When a short screening IQ test is being developed, it should establish concurrent validity evidence by showing highly correlated results with either the Stanford-Binet or WISC-III.

Predictive validity evidence refers to how well a person's result predicts future behaviours or accomplishments. This type of validity evidence is particularly important and useful for tests such as college or professional school entrance tests and aptitude tests.

Although criterion-related validity is well accepted, an appropriate method to measure the criterion variable is a concern. For example, in the area of sports physiology, a maximal oxygen consumption test (VO_{2max} test) is acknowledged as a fitness test for aerobic endurance. To decide that one has reached VO_{2max}, several difficult criteria, which many individuals cannot meet because of exhaustion, should be met. However, in adapted physical activity, many individuals with disability have difficulty reaching their true maximal VO_2, thus researchers often use a peak VO_2 test as the criterion for measuring aerobic endurance (Pitetti, Jongmans, & Fernhall, 1999).

Construct validity. Construct validity is the collection of empirical evidence to support the existence of the theoretical construct that underlies measurement and the resulting inferences (Burton & Miller, 1998). Construct validity deals with whether the assessment tool adequately detect the assessor's abstract idea to assess (Kubiszyn & Borich, 2003). Thus, the fundamental issue of construct validity is how we make appropriate inferences to what we cannot directly observe (Bouffard, 2003; Salvia & Ysseldyke, 2004). The unobservable ideas must be converted (constructed) in the words

of human language or graphics to understand. Thus, construct validity is normally demonstrated with statistical methods that show whether or not a common factor can be shown to exist underlying several measurements using different observable indicators (Chatterji, 2003).

To obtain construct validation, the construct should be defined since constructs are implicit variables. This is simliar to providing content-related evidence which defines universe or domain of content (Yun & Ulrich, 2002). Using appropriate statistics and rationale to test the hypotheses is the next step to obtain construct validation. Hypotheses are related to inferences, and biased items that systematiclly influence test scores could distort the hypotheses (Yun & Ulrich, 2002). Although there is no absolute method to selecting the most appropriate instrument, it is important to carefully examine the possible item bias and intended purpose of measurement. The last step of obtaining construct validation is gathering empirical evidence to support the hypothsis (Yun & Ulrich, 2002). Correlations can also be used in establishing construct validity. For example, a high correlation coefficient should result when measuring similar variables (verbal performance and reading performance) with two different assessment tools. Simultaneously, the score of a reading test should not correlate with a measure of a different construct, such as physical strength.

Reliability

Reliability implies stable, consistent and predictable outcome from one use(r) to another, under different conditions (Chatterji, 2003; Kubiszyn & Borich, 2003). Four methods of estimating reliability are test-retest method, alternate form techniques, split-half method, and internal consistency analysis.

Test-retest method. This method confirms the temporal stability of the test. This method deals with consistency of results among different testing occasions. In other words, with the same test tool in the same context and to the same person or groups of people, the test is given twice and the correlation between the first set of scores and the second set of scores is determined. The time period between first test and second test could affect on the result of the second test. The longer time has given, the more the characteristics of the study groups could be changed.

Alternate form techniques. While the test-retest method determines consistency of results among different testing occasions, alternate form techniques engage the consistency among two or more different forms of a test. If there are two equivalent forms of a test, these two forms can be used to obtain as estimate of the reliability of the test. This method eliminates the time gap problem in the test-retest method. To use this method of estimating reliability, two equivalent forms of the test must be available, and they must be administered under conditions as nearly equivalent as possible (Kubiszyn & Borich, 2003). However, obtaining equivalence of the two different tests is difficult. This may result in ambiguity of interpreting the reliability estimate; the tool may be unreliable or the equivalence of the two assessment tools was not done effectively.

Split-half method. This method divides the test into two separated tools. Then, the total score for each student on each half is determined and the correlation between the two scores for both halves is computed. It does not require two tests, and the time gap problem is not involved. Therefore, this method is most frequently used for estimating the reliability of classroom tests (Kubiszyn & Borich, 2003). To use optimally, homogeneity of all items of the assessment tool should be obtained, and the number of the items should be large.

Internal consistency analysis. It is believed that person who gets one item correct will likely get other, similar item, correct. In other words, items should be correlated with each other, so the test ought to be internally consistent. One test could be split into two separated tests in accordance with the correlation of each item. This can be done by assigning all items in the first half of the test to one form and all items in the second half of the test to the other form. To use this method, all items of varying difficulty should be randomly spread across the test.

A good assessment is valid and reliable. Note that reliability sets the upper limit of a test's validity. Thus, an assessment may be reliable but invalid or unreliable and invalid, but an assessment can not be unreliable and valid (Salvia & Ysseldyke, 2004).

Objectivity

The objectivity of an assessment refers to the degree to which equally competent raters obtain the same results (Linn & Gronlund, 2000; Ward & Murray-Ward, 1999). When implementing classroom assessment constructed by teachers or performance-based assessments by schools, objectivity may play an important role in obtaining reliable measures of achievement. In a performance-based test requiring judgemental scoring, the results depend on the person doing the scoring. Different persons get different results, and even the same person may get different results at different times. Using objective tests only or abandoning all methods of assessment that require judgmental scoring in order to enhance the degree of reliability would have an adverse effect on validity. A better solution is to select assessment procedures most appropriate for the learning goals being assessed and then make the assessment procedure

as objective as possible (Linn & Gronlund, 2000). For example, in the use of essay tests, objectivity can be increased by careful phrasing of the questions and by a standard set of rules for scoring system so that different raters produce the same or very similar scores. Careful training of raters to use the scoring system is also required. Such increased objectivity will contribute to better reliability without sacrificing validity.

Philosophical background of authentic assessment

Constructivism

Based on the work of Dewey, Piaget, and Vygotsky, constructivism views children as actively interpreting their experiences in physical and social environments and thus constructing their own knowledge, intelligence, and morality. Thus, teachers do not just transfer knowledge to learners but individual learners connect what teachers expect them to learn with their own experience. Learners make personal meaning for themselves, develop shared meaning with others, and then reflect on their meaning in the public arena of the classroom (DeVries, Edmiaston, Zan, & Hildebrandt, 2002; Fosnot, 1996; Gabler & Schroeder, 2003; Gagnon & Collay, 2001). In his declaration "My Pedagogic Creed," Dewey (1897) placed great emphasis on the broadening of intellect, development of problem solving skills, and critical thinking skills, rather than the simple memorization of lessons. With such a viewpoint, he focused on student capacity, interest, and habit through the establishment of interactive, student-centered "learning communities" within the classroom (Dewey, 1934).

Piaget believed that the human was a developing organism, not only in a physical, biological sense, but also in a cognitive sense. Rather than adopting ideas of behaviourism or maturationism, concept development and deep understanding were

considered, and learning stages were understood as construction of an active learner. Investigating the influence of experience on children learning in various settings, Piaget argued that the learning process was dynamic, multidimensional, and nonlinear, and it could be reconstructed with existing knowledge into new cognitive structures (Fosnot, 1996).

While Piaget's work focused on the cognitive structure of individual, Vygotsky emphasized the importance of a social context as learners actively construct knowledge. Vygotsky (1978) asserted that teacher to student transmission of knowledge is not the only efficient way of obtaining knowledge. Rather, the student should be immersed in an interactive setting, where learners have some degree of control over the nature and reciprocal learning activities, facilitating new cognitive structures within learners. From constructivism, learning is best promoted through an active process emphasizing interaction and the use of knowledge of real situations. Gabler and Schroeder (2003) have suggested several important points teachers need to recognize from perspectives of Dewey, Piaget and Vygotsky:

1. Learners of any age should make practical application of new experiences by relating them to their previous experiences. Making ideas understandable from a learner's point of view is essential for deep learning.

2. Although it may be necessary to memorize certain facts as part of a learning experience, deeper learning involves cognitive restructuring on the part of the student. The teacher's most important role is to facilitate active learning process.

3. Learning is something that a learner does, not something that is done to the learner. Students must be involved in the learning process, making their own inferences and experience. Students need to struggle with new ideas that contradict or differ from

existing knowledge.

4. Effective teaching involves continual probing of the nature of student understanding.

5. Deeper understanding includes gaining insight into the connections between disciplines and knowledge of the ways of thinking within them.

6. Superficial information is the result of teaching that emphasizes covering content rather than building student understanding through active student experiences both within the classroom and in the world at large.

7. Reflection (i.e., thinking carefully about what we are doing and why) is a vital part of effective teaching, which promotes the learning of students and the empowerment of teachers as professionals.

Multiple intelligences

The traditional intelligent test, created by Binet in 1904, has been used and developed to measure human intelligence around the world. In the 1980s an American psychologist, Gardner (1993), pointed out a limitation of the traditional method that measures human intelligence in linguistic, logical/mathematical areas only. Alternatively, he suggested multiple intelligences. Gardner's multiple intelligence theory disagrees with putting human intelligent into a single grid, such as IQ tests, and questions whether human intelligence can be measured by standardized pencil-and-paper tests, such as multiple choice tests.

Gardner insisted that five other categories, visual-spatial, body-kinesthetic, auditory-musical, interpersonal communication, and intrapersonal communication, should be added to traditional intelligent tests, which have been limited to linguistic and

logical/mathematical categories. Knowledge is explained in multiple intelligences theory as follow:

First, knowledge is constructed of different and divided intelligences (Language, logic, mathematics, space, music, movement, and interpersonal /intrapersonal communication). Individuals inherently possess ability that develops the seven intelligent areas. However, while it hypothesizes all individuals will develop each intelligence area; they will not have the same intelligent profile, and intelligences forms different types as cultural differences.

Second, intelligences are independent of each other. In other words, a measurement in a certain area is not able to predict performance in other areas. Humans have many different abilities, but ability of one field is not related to other fields. A genius is limited in a special area.

Third, intelligences act reciprocally. While each intelligence is independent, they do act altogether. For instance, when solving a math question, linguistic, logic, and mathematic abilities should work in a reciprocal manner to solve the question.

Multiple intelligences are found in every person but at different levels, and education and training seem to promote the intelligences. These intelligences act in combination to accomplish a purposeful activity. For example, to cook with a recipe requires linguistic intelligence for understanding the words, logic and mathematic intelligence for organizing the levels of cooking process, body-kinesthetic ability to control the hand movements, and interpersonal intelligence to meet the tastes of each member of the family. Thus, an aptitude occurs when several intelligences play altogether.

The emphasis of constructivist discourse in education is that students should

create their own educational environments to make personal meanings, in which they share knowledge and intelligence with peers. In accordance with Gardner's ideas, students who are complex should be taught and assessed by many different approaches. In order to achieve these objectives, and to complement weaknesses of traditional assessment, an alternative approach should be introduced in education for assessing students.

Authentic assessment and rubrics

Definition and characteristics of authentic assessment

The definition of authentic assessment requires clarification of other terms, such as "traditional assessment" and "alternative assessment." Traditional assessment is commonly used to refer to standardized tests, norm-referenced tests, and multiple-choice tests. More specifically, a traditional test can be thought of as the traditional popular, structured-response, written test (Chatterji, 2003). "Alternative assessment" distinguishes itself from "traditional" in its open-ended response format. This new approach includes new types of test, such as interview and performance test that resemble real-life settings (Wiggins, 1998). Because of the open-ended format, most alternative assessments call for the use of fixed clear standards to avoid judgments and observations being based on personal biases or errors from the tester (Chatterji, 2003).

The term "authentic assessment" is not as easily clarified as the first two, since it enjoys multiple and varied usages. In efforts to overcome the limits of traditional assessment, Wiggins (1989) advocated that authentic assessments were tools for raising standards and expectations of schools. Teachers are dissatisfied with traditional forms of assessment. Teachers know that students are learning, but the tests do not reflect that

learning nor do they facilitate learning. As part of this process, students often perform tasks that have no worthwhile or real-life counterpart. In physical education, for example, tests of motor ability, fitness, sports skills, knowledge, and psychosocial traits may be objective and reliable, but they may also fail to measure actual outcomes or objectives of interest to the teacher and students (NASPE, 1995). Wiggins argued that school tests should be changed to a more practical and realistic format. Changing the purpose of a test greatly influences the instruction, curriculum, and assessment (Wiggins, 1989). Wiggins (1989; 1998) promoted authentic assessment as a method designed to improve performance of students by showing them the highest level of performance that they are capable of achieving. Moreover, it simultaneously requires performances and presentation of meaningful tasks conducted in the real world. For example, to assess a surgeon, anatomy knowledge should be assessed, but also how well he/she performs a real operation. Dancers should dance, singers should sing, and athletes should perform.

Authentic assessment should require higher-level thinking. It does not require students to only write down what they know. Authentic assessment measures the ability to think and to apply general concepts to a variety of situations (Block et al., 1998). Authentic assessment redefines the meaning of understanding as employing knowledge wisely, fluently, flexibly, and aptly in particular and diverse contexts (Wiggins, 1993). Therefore, to assess learners understanding and status more accurately, several performances in various contexts should be required. Lund (1997) gives a good example of a fitness class. Students learn the components of fitness as well as how to analyze their current fitness status. Using this information, they are required to synthesis and apply the knowledge and are able to create an exercise program that will achieve fitness goals for their future life. Developing higher-level thinking helps

students achieve more complicated learning.

Authentic assessment should set a very clear standard. Since personal observations and opinions of the assessor are clearly involved, assessment standards should be established. One of the most significant aspects of authentic assessment is that students no longer ask teachers such questions as "Is this what you want?", or "Will this be on the test?" Importantly, clear goals and standards should be introduced in advance so that students know what is expected of them and how they will be assessed accordingly.

Authentic assessment should be embedded in the curriculum whereby it enhances student learning, not just documents their learning status. Since authentic assessment is directly linked to the curriculum and instruction, teachers and students can make adjustments during the unit. Students are offered many opportunities to practice, rehearse, consult, and to receive feedback to refine performances and productions (Wiggins, 1998), which follows with an expectation for them to produce high quality performances. In other words, authentic assessment acknowledges that both the product and the process of learning are important.

In summary, authentic assessment is an alternative assessment to traditional assessment. The main purpose of authentic assessment is to improve student performance in real world settings. It is the comprehensive assessment system that assesses both process and products of self-meaningful performances by students in a direct, continuous, and holistic manner.

Rubrics

Rubrics are one of the authentic assessment tools to help both teacher and student

to assess critically. The original meaning of word "rubric" is from the Latin word *ruber* meaning "red." In literacy history of mid-15th century, Christian monks invariably wrote a large red letter when they reproduced each major section of sacred books. Because of the Latin word for red, rubric became headings for major divisions of books (Jackson & Larkin, 2002; Popham, 1997; Schultz, 2002).

In current usage, *rubric* began to take on a new meaning among educators. A rubric is a carefully described scoring tool to judge student performance, as well as guidelines for students to focus on key elements to achieve the mastery of performances (Goodrich, 1997; Lund, 1997; Montgomery, 2000; Popham, 1997; Smith & Cestaro, 1998). This new assessment system has gained popularity in education for many purported benefits.

Benefits of using rubrics. Rubrics appeal to both teachers and students for many reasons. First, students can understand the qualities associated with a specific task or assignment (Block et al., 1998; Whittaker, Salend, & Duhaney, 2001). Since rubrics show the hierarchy of the specific tasks from low to high level performance, and explain every detail of each step to achieve the quality performance, students clearly recognize teacher's expectation, as well as they become more aware of their own personal strengths and weaknesses (Hall & Salmon, 2003).

Second, rubrics help students to be more self-determined in their learning (Block et al., 1998). Rubrics indicate what skills a student currently has and what other skills the student needs to upgrade to the next levels. In such a way, students assess themselves, which makes them active learners, and increases their sense of responsibility for their own work. By self-assessment, teachers guide students toward making realistic goals for improvement (Montgomery, 2000). Teacher awareness of the student's ability

to self-assess accurately provides valuable information how deeply the student understands the task (Montgomery, 2000). Moreover, students are able to assess their peers with the same rubric, which is a very significant function of rubrics for students, especially those with disabilities. In general, students with disabilities have been recognized as passive learners. In many cases, they receive services from others, such as teachers, caregivers, or peers. This unequal-status relationship could make the students more passive learners. However, rubrics could be a good mediator amongst students. In a case of peer tutoring setting, a student with a disability and the able-bodied partner can practice together with a rubric that enables the student with a disability to potentially guide the student without a disability. Such reciprocal teaching could create equal-status in integration/inclusion education settings (Sherrill et al., 1994).

Third, from the teacher's perspective, they enjoy using rubrics for both teaching and assessment (Goodrich, 1997; Whittaker et al., 2001). Teachers use rubrics for assessing student performances, products, and progress, where teachers can give feedback to students for improving their work. Since rubrics are directly embedded in instruction and can be expanded and specialized, teachers can utilize rubrics to create individualized education programs (IEP), which delineate individual student short and long term goals in heterogeneous classes (Block et al., 1998; Goodrich, 1997).

Fourth, teachers can clarify and communicate their expectations (Montgomery, 2002). Teachers can deliver their expectations to students with rubrics, and rubrics can help teachers explain their grading of student work to family members. Since family members also can use rubrics to assist their children with assignments, family members can be involved in the learning process (Whittaker et al., 2001).

Types of rubrics. Scores are awarded based on predetermined criteria set forth in

the rubrics. Depending on the type of rubric used, grades are awarded by the total score only (i.e., holistic) or by separate pieces being judged and then reached at a final score (i.e., analytic). Holistic rubrics are more product-oriented. Once the final product is submitted, summative assessment with a holistic rubric is accomplished as the rubric is used to award a final grade. A holistic rubric rates an activity in its entirely without regard to the separate pieces. This type of rubric is used when the components of an activity are too interrelated for easy division (Jackson & Larkin, 2002). Analytic rubrics are more process-oriented. By distributing ahead of time what is expected and using the criteria as expectations, the rubric guides students throughout the unit of instruction. In this way, progress toward a goal and the process of learning is evaluated; therefore, formative assessment is accomplished. An analytic rubric rates separate pieces of an activity individually and then add all scores for a total rating. The rubric can be shared among teachers, students, and parents to clarify and refine the criteria toward the final product (Jackson & Larkin, 2002).

Create and use a rubric. In various Internet sites and books, many samples of rubrics are provided and ready to be used. However, every education context is unique, and there is no rubric that fits all contexts. Thus, it is best to design a rubric for each class and each student. When creating a rubric, one needs to remember that rubrics represent not only scoring tools but also, more importantly, instructional guidelines. Thus each level of a rubric must represent a key attribute of the skill to teach as well as being assessed (Popham, 1997). Each criterion must be teachable in the sense that teachers can help students increase their ability to use the criterion when actually required.

The following steps are suggested when designing and using rubrics. In the first step, teachers decide on the final end product(s). If the teacher collaborates with

students, some examples of poor to exemplary work are shown to students, then they identify the characteristics that classifies the levels (Goodrich, 1997). Woods and Anderson (2002) asked 19 students in grade 9 to design and implement a rubric in an aerobic unit. They reported that most students agreed on the scores awarded to the performance of peers, and students were much more satisfied with rubrics they had designed than those of teachers. Some negotiation would be required to develop criteria that both students and teachers believed was important to arrive at an acceptable rubric format (Woods & Anderson, 2002).

Second, criteria and weight are determined (Goodrich, 1997; Hall & Salmon, 2003; Lunsford & Melear, 2004). Rubric designer(s) discuss the qualities of an exemplary response and all essential components of the desired performance. At this point, they make a checklist of all criteria, and articulate gradations of quality. Best and worst levels of quality should be established, and then middle levels based on the knowledge of common problems are filled in between the best and worst levels.

Third, decide who will assign the grade (Lunsford & Melear, 2004). Teachers traditionally assume full responsibility for grading. However, some may wish to share the responsibility of grading with students (self/peer) and expert judges. Involving students in the evaluation process gives them a deep level of ownership and heightens their interest in the task being assessed (Lunsford & Melear, 2004; Montgomery, 2000). Sometimes experts may be called on to evaluate students from essay contests in elementary school to high school science fairs. Expert judging often takes place at the highest levels of academia such as evaluating a master student's thesis. Even professional research papers are reviewed by experts before publication, where experts may be given a checklist or rubric to guide their work.

Step Four is the use of a rubric to receive feedback about effectiveness from the users. Let students use the rubric to assess the models used in the first step. If there are interpretation problems, ask the students to ask clarification questions and make comments as they evaluate (Montgomery, 2000). Fifth, use the questions and comments to evaluate and revise rubrics. It is important to examine the rubric's effect on students, teachers, and other relevant parties. Information from students as well as teachers, and family members can be helpful in examining the overall effectiveness of the rubric. For example, students and family members can provide information about how the rubric aided or hindered performance. Similarly, teachers can observe how the rubric reflects the teaching and learning process (Whittaker et al., 2001).

Once a rubric is created, teachers need to teach and encourage students to use the rubric. Before starting the actual assignment, it is helpful to give the students the created rubric and use it to evaluate several sample assignments of varying quality (Stutzman & Race, 2004; Whittaker et al., 2001). This enables the students to solidify those indicators they include in the actual assignment and also to determine if the rubric is workable.

In summary, authentic assessment is a new approach to enhance educational environments where the teacher directly assess what students know and how they are learning, and at the same time, students clearly understand the meanings of their learning by active participation in the learning process. Authentic assessment can be an ideal solution to develop educational status by high correlation between what students learn at schools and what they really do in the real world. As a tool of authentic assessment, rubrics facilitate students to be more self-determined in their learning. With self- and peer-assessment with rubrics, students have ownership of their learning. Especially, it

encourages students with a disability to assess their peers that create reciprocal equal-status environment in integration/inclusion education settings (Sherrill et al., 1994).

Chapter 3

METHOD

The purpose of this study was to investigate the psychometric properties of rubrics in an integrated swimming class. Swimming can improve all five components of physical fitness: aerobic fitness, muscular strength, muscular endurance, flexibility, and body composition (Sova, 1995). Also, water buoyancy offers a friendly and safe environment for physical activity for all people, including young children and those with disabilities. Therefore, teaching swimming is sound educational practice. This chapter is divided into the following sections (1) Participants (2) Development of rubrics (3) Task (4) Procedure, and (5) Data analysis.

Participants

There were four phases to the research. The first phase established the participants who were able to use the rubrics. Only students who attained 80% efficiency were selected for the subsequent three phases and data analyses.

Twenty-six children (9 boys and 17 girls) between the ages of 8 and 13 years returned consent forms and therefore were included in phase 1. They attended Mackay Centre School (http://www.emsb.qc.ca/mackay/) in the Montreal area. Mackay Centre School focuses on rehabilitation while the school concerns itself with the educational needs of the children. It is equipped with a heated swimming pool, a library, a computer room, a gymnasium and an occupational therapy room. A special point of Mackay Centre School is a "reverse integration" program; children without disabilities attended the school which had full therapeutic services for the students with physical and developmental disabilities. In this manner, it was hoped that social integration of same-

aged peers with and without disabilities could emerge. Approximately two-thirds of the children at the school had a disability. Both the children with and without disabilities participated in classes together. The physical education program in the school was fully inclusive and consisted of aquatics and physical activities in the gym. Swimming classes were taught by the physical education teacher, once a week for approximately 45 minutes. All participants were recruited from one of four classes, where the size of each class ranged from 9 to 11 students. There were two groups of students who returned consent forms: 14 with physical disabilities (e.g., cerebral palsy, developmental delay, and others) and 12 able bodied children. Table 1 provides descriptors of these 26 children. Nineteen students (5 boys and 14 girls) qualified to participate in the phase 2 and beyond because they achieved 80% efficiency in using the rubrics (see Procedures below). Table 2 shows the descriptors of these 19 children.

In the phase 2, the swimming of 20 meter, two from 19 qualified students did not participate, because they were absent. The seven students who were disqualified continued to participate in the study in various ways. Four students joined the second phase, swimming 20 meters. Thus, 21 students (17 who reached 80% plus these four) participated in the swimming which was videotaped for subsequent analyses. Overall, those who were disqualified did not know they were eliminated and participated until phase 4, but their data were not used for analyses. This strategy prevented students from being frustrated by elimination from the study.

Three teachers (physical education teacher, main researcher, and a research assistant) were included in the study. The physical education teacher had taught physical education and swimming classes for 7 years. He was well aware of the context of disability. The main researcher had worked for the school for three years as a

physical education assistant both in the gym and the swimming pool. Consequently, his work experience enabled him to implement this research with the young participants in their natural school setting. The research assistant was a female master's student of adapted physical activity. She had been life guarding and teaching the Red Cross program for 10 years.

Table 1. Descriptors of the 26 children

Descriptor	N (%) in total sample
Gender	
Male	9 (35)
Female	17 (65)
Grade & Age	
3	6 (23) – 8.11 to 9.6 years
4	9 (35) – 9.9 to 12.9 years
5	3 (11) – 11 to 11.4 years
6	8 (31) – 11.10 to 13.3 years
Disability	
Without disability	12 (46)
With disability	14 (54)
Disability Type	
Cerebral Palsy	3 (21)
Developmental delay	4 (29)
Other	7 (50)

Table 2. Descriptors of the 19 children

Descriptor	N (%) in total sample
Gender	
Male	5 (26)
Female	14 (74)
Grade & Age	
3	5 (26) – 8.11 to 9.6 years
4	7 (37) – 9.9 to 12.9 years
5	3 (16) – 11 to 11.4 years
6	4 (21) – 11.10 to 12.11 years
Disability	
Without disability	12 (63)
With disability	7 (37)
Disability Type	
Cerebral Palsy	2 (29)
Developmental delay	2 (29)
Other	3 (42)

Development of rubrics

Rubrics are widely used in education and various rubrics have been created and modified into specific context by teachers. Rubrics are purported to be effective assessment tools in adapted physical education in order to include students with disabilities into regular physical education. Lieberman and Houston-Wilson (2002)

presented many such rubrics that covered a host of physical activities. They argued those rubrics are helpful for teachers to critically assess students, and for students to enhance skills by understanding criteria for mastery. Also, rubrics can be developed in ways which individualize instruction for students.

Two swimming rubrics were created, each included 10 levels each (see Appendix D & E). To enhance ecological validity, the rubrics were designed to be implemented in the regular swimming classes of the school. To teach swimming and assess students, the physical education teacher used Aqua Quest, the standards for swimming set forth by the Canadian Red Cross Society (1996). This swimming program is designed to enhance swimming skills and water safety for ages 4 to 16 years. Aqua Quest is composed of levels 1 to 12, the highest being 12. In order to encourage the students to practice and develop their swimming skills, the rubrics were created from Aqua Quest performance criteria. Because the main purpose of the study focused on front crawl, the first three levels of Aqua Quest were not considered in developing the rubrics, since those levels deal with water orientation and floating rather than the front crawl per se. Since the physical educator encouraged the students to reach the highest level, the performance criteria for the Shark, the highest level of the rubrics, was matched to Aqua Quest level 12. Rubric A was designed for children who swim 20m (two widths of the school swimming pool) front crawl without a floatation device, while rubric B was created for children using floatation devices, such as life jackets, a swimming noodle, and/or a swimming belt. In the case that the student believed more support was needed to swim independently, two or more of the devices could be used at the same time. In order for the young participants to avoid discomfort in being assessed, and to enhance enjoyment, each level of the rubrics was identified by the name of a sea creature. Using words

rather than numbers to represent levels helps students to move away from linear percent scales (Stutzman & Race, 2004). The following are the sea creatures used in the levels of rubrics: Starfish (1), Sea urchin (2), Jellyfish (3), Shrimp (4), Seahorse (5), Goldfish (6), Blowfish (7), Tuna (8), Dolphin (9), and Shark (10). Each level of the rubrics was comprised of four parts: Arms, Feet, Head/Body, and Breathing. It helped students judging performances critically. Importantly, the two experts were the physical education teacher who had seven-years of experience of teaching swimming at the school and the research assistant who had two years of experience at the school and 10 years in teaching regular swimming classes. These individuals reviewed the rubrics several times and agreed that they were suitable for this study.

Task

All participants were asked to swim front crawl for 20 meters. The size of the swimming pool at the school was 25 X 10 meters. Thus, participants were requested to swim the width of the pool and return. Those who wanted to perform with one or several floatation devices (e.g., lifejacket, swimming noodle, or swimming belt) were given the opportunity to select the floatation device of their choice. All performances were videotaped. A video camera (Sony DCR-TRV17) was supported on a tripod, and managed by the main researcher. The camera was placed at the end of the turning point, slightly off to the side to capture the front, back and side view of the swimmer. The video was presented on a TV monitor in a subsequent class for teachers, peers and swimmers to evaluate each performance.

Procedures

There were four phases to the procedures; (1) rubric teaching, (2) swimming and self-assessment, (3) video assessment, and (4) temporal stability video assessment. The rubric teaching phase was held one week prior to the swimming and self-assessment phase. The main researcher and research assistant taught the participants how to use the rubrics. All participants (N=26) were asked to come to a room during the lunch hour. Paper copies of rubrics A and B were given to each participant. The main researcher and the research assistant explained every detail of the performance criteria of the two rubrics, with a prepared rubric teaching videotape showing the 10 levels of each rubric. Students were encouraged to ask questions during the process if there was confusion. The main researcher explained by demonstrating and/or rephrasing the instruction in a way that was more easily understood. At the end of these demonstrations and explanations, the researcher asked if anyone had additional questions. When no further questions arose, the researcher concluded that all students understood the details of the criteria of the rubrics.

At this point, competency in using the rubrics was determined. Five levels from each rubric were showed on the TV screen. The selected performances (rubric A-Sea Urchin, Goldfish, Jellyfish, Tuna, and Shark; rubric B-Jellyfish, Seahorse, Dolphin, Goldfish, and Sea Urchin) were presented in mixed order for the testing video tape. Students assessed the sample performances on the screen using both rubrics A and B. After completing the test, the researcher evaluated the assessments made by the students on the sample performance video. Seventeen students (12 without disability, 5 with disability) scored 80% (8 out of 10) or higher, the criterion established to take part in phase 2, 3, and 4. Those who scored below 80% (9 with disability) were given another

rubric teaching session the following day at the same time. The same procedures took place as the first day. Seven students once again scored below 80% and were disqualified as participants in this study. Consequently, 19 students (5 boys and 14 girls) were qualified to continue as full participants. Table 2 shows the descriptors of 19 children.

One week after phase 1, the 17 students swam 20m front crawl as well as four others who had not met the criterion. Thus, for phase 2, 21 were videotaped. In order to facilitate participation, the regular swimming class times were used to videotape the swimming, because physically assistance (e.g., changing, toileting, and transferring) was readily available to students with disabilities at that time. During the swimming class, the participants were asked to perform front crawl/front swim across the 2-widths of the pool at a depth of chest level. Chest level was chosen so that the swimmers could feel comfortable and safe in the water (CRCS, 1996). Immediately, after performing the 20m front crawl swimming, each participant was asked to self-assess. Of course, only the data for the 19 qualified students were included in the analyses.

The third phase started the next week. Video footage had been transferred to VHS to play on the TV. The students from each class were asked to come to the research room to assess the performances of the participants from their class. This procedure was implemented for each class. However, two older classes (grade 5 and 6) were merged as one group, because of low numbers, and because these classes were often joined in other school activities. Three teachers (physical education teacher, main researcher, and the research assistant) and participants, sat before a 56-inch TV screen. Each person had rubrics A and B and assessed what they saw on the screen. The three teachers assessed 21 students, 21 measures were obtained as peer assessments, and 17

students self-assessed their own performances. The reason why only 17 students were included for self-assessment was two students of the 19 could not participated in the swimming performance on the day of self assessment. Assessors were not allowed to talk to each other during the assessment time. Once assessment was complete, all assessors handed in the assessment rubrics to the main researcher and left the room.

To determine reliability, the fourth phase was implemented. Participants were asked to return to the research room one week following the first video assessment. The same procedures took place as the previous week. These data permitted assessment of temporal stability. Appendix F and G provide all video assessment data.

Data analysis

The rubric data can be ranked from 1 to 10; starfish to shark. Since the data are ordinal ranks , non-parametric tests were used for analyses (Daniel, 1978; Field, 2000). Statistical program SPSS 13.0 for Windows (SPSS Inc.®, Chicago, IL) was used.

Teacher, peers, and self

In order to obtain an individual score for each participant from teachers, the median values from the assessment of the three teachers for each student were used. Median values represents better central tendency with ordinal data, because they are not affected by the size of any extremely large or extremely small values (Welkowitz, Ewen, & Cohen, 2002). Similarly, the median values from students in the same class constituted the score of each individual for peer assessment analyses. Since each student evaluated him/herself only once, that score was used as the value for self-assessment analyses.

Spearman's correlation coefficients were run to determine temporal stability of teacher, peers, and self assessments. The Kruskal-Wallis Test, non-parametric test for more than two independents variables, was used to compare differences of the three groups (Teacher vs. Peers vs. Self). A significant Kruskal-Wallis test was followed by three Mann-Whitney tests (Teacher vs. Peers, Teacher vs. Self, and Peers vs. Self).

Students with and without disability - peer and self-assessment

To determine relations between participants with and without disability, Spearman's correlation coefficients were calculated for temporal stability for each group. The Mann-Whitney test was used to compare students with and without disability. The first analysis was based on peer assessment of 21 participants. This would provide evidence of equitable assessment by the two groups of peers. The second Mann-Whitney analysis compared teacher and self assessment for each group. They would permit inferences about equitable assessment by the student and the teacher.

Peer and self assessment - male and female students

Analysis of gender relations was completed with the same procedures as disability. After determine the reliability with Spearman's correlation coefficient, the Mann-Whitney test was used to compare: male and female students for peer and self assessment.

Immediate assessment and video assessment

Additionally, one more comparison was analysed; video self assessments were compared with assessments done immediately following swimming. While video

assessment has some advantages, it was considered important to compare video assessment with immediately self assessment, since judging the video may not to be completely authentic. A Spearmen's correlation coefficient was calculated comparing immediate self assessment with first week video self assessment.

Chapter 4

RESULTS

The purpose of this study was to investigate the psychometric properties of rubrics in an integrated swimming class. This chapter is divided into four major categories (1) Teacher, peers, and self assessment comparison (2) Students with and without disability (3) Male and female students (4) Immediate assessment and video assessment.

Teacher, peers and self assessment comparison

The test-retest temporal stability was significant ($p < 0.05$) as showed by the coefficients of correlation: $r = 0.92$, 0.83, and 0.70 for teachers, peers, and self evaluation, respectively. Table 3 provides descriptive data of three groups for two weeks. Due to the significant temporal stability, the data of week 1 were used for further analyses. Overall, each of the groups was consistent in using the rubrics over the two weeks.

Table 3. Descriptive statistics of three different assessors for weeks 1 and 2

	Week 1			Week 2		
Assessors	N	* Mean	** SD	N	Mean	SD
Teachers	21	7.38	1.96	21	7.52	1.91
Peers	21	6.86	1.44	21	7.10	1.35
Self	17	7.35	1.80	17	7.35	2.15

* Mean - mean values of each median value.

** SD - Standard deviation.

The Kruskal-Wallis test comparing the results obtained from teachers, peers, and self assessment (see Table 3) showed that the three groups were not different when assessing performance ($H(2) = 2.83, p > 0.05$).

Students with and without disability

Table 4 provides information regarding temporal stability between weeks 1 and 2 for peer and self assessment for students with and without disabilities. There were statistically significant correlations for three of the groups ($p > 0.05$). However, self assessment for students with disability ($r = -0.30$) showed a low and non-significant negative correlation. For this reason, further statistical comparisons from this group were made separately for weeks 1 and 2.

Table 4. Spearman's Correlation coefficients from peers and self assessment of students with and without disability

Peer	
Without disability (N=12)	$r = 0.80; p = 0.00$
With disability (N=7)	$r = 0.69; p = 0.00$
Self	
Without disability (N=11)	$r = 0.91; p = 0.00$
With disability (N=6)	**$r = -0.30; p = 0.95$**

The Mann-Whitney comparison of peer assessment demonstrated that students with disability (*M=6.4, ±1.43*) produced lower scores than students without disabilities (*M=7.17, ±1.37*) ($U = 142.5, p < 0.05$). However, the same comparison of the second week's assessment showed that peer assessments of the two groups were not significantly different ($U = 168.5, p > 0.05$). When comparing teacher assessment to self assessment, there were no differences for the students with a disability ($U = 13.0, p > 0.05$) or without

disability ($U = 51.0, p > 0.05$). Thus both groups gave themselves similar scores as did their teachers. This was also true for students with disabilities in week 2 ($U = 14.0, p > 0.05$). Table 5 provides descriptive data of peer and self assessment for students with and without a physical disability.

Table 5. Descriptive statistics of peer and self assessments for students with and without disability, weeks 1 and 2.

Type	Peer (***W1)		Peer (W2)		Self			
Assessors	* SWOD	** SWD	SWOD	SWD	SWOD	Teacher	SWD	Teacher
N	21	21	21	21	11	11	6	6
Mean	7.17	6.40	7.14	6.60	8.45	8.64	5.33	6.17
SD	1.37	1.43	1.58	1.34	0.93	0.67	1.03	1.38

* SWOD, Students with out a disability

** SWD, Students with a disability

*** W, Week 1 and 2

Male and female students

Table 6 shows the reliability between weeks 1 and 2 regarding peer and self assessment for male and female students. There were statistically significant correlations for three of the groups ($p > 0.05$). However, self assessment for male students ($r = 0.32$) showed a low and non-significant correlation. For this reason, statistical comparisons from the male student were made separately for weeks 1 and 2.

Table 7 provides descriptive data of peer and self assessment for male and female students. Boys did not differ in assessing swimming performance of peers from girls ($U = 175.5, p > 0.05$). For self assessment, boys ($U = 4.5, p > 0.05$), ($U = 5.0, p > 0.05$, for

session 2) and girls ($U = 69.5$, $p > 0.05$) showed similar competence in comparison with their teacher. Thus, there were no gender differences in assessing rubrics scores to self or peers.

Table 6. Spearman's Correlation coefficients from peers and self assessment of male and female students

Peer	
Male	$r = 0.79$; $p = 0.00$
Female	$r = 0.82$; $p = 0.00$
Self	
Male	**$r = 0.32$; $p = 0.68$**
Female	$r = 0.61$; $p = 0.03$

Table 7. Descriptive statistics of peer and self assessment for male and female students

Type	Peer		Self			
Assessors	Male	Female	Male	Teacher	Female	Teacher
N	21	21	4	4	13	13
Mean	6.31	6.76	6.50	7.50	7.62	7.85
SD	1.33	1.72	2.38	1.29	1.61	1.82

Immediate assessment and video assessment

Table 8 shows the descriptive data of video and non-video self assessments. Self assessment provided immediately after each individual's swimming without the video-recorded feedback was significantly correlated with video self assessment ($r = 0.76$; $p < 0.05$).

Table 8. Descriptive statistics of immediate assessment and video assessment

Type	Video	Non-Video
N	17	17
Mean	7.35	7.53
SD	1.80	1.62

Chapter 5

DISCUSSION AND CONCLUSIONS

The purpose of this study was to investigate the psychometric properties of rubrics in an integrated swimming class. This chapter includes four major categories (1) Validity (2) Reliability (3) Objectivity (4) Conclusions, and (5) Recommendations for future study.

Validity

Authentic assessment is focused on the context where students actually perform (Wiggins, 1989, 1998). Teachers want to know how successfully students perform in the real world outside of school settings. Thus, authentic assessment looks like it is measuring the desired behaviour. It is consistent with the idea of face validity. However, face validity is not enough evidence to say that an assessment is valid (Joyner & McManis, 1997; Yun & Ulrich, 2002). Joyner and McManis (1997) argued that quality of authentic assessment is determined by three components: validity, inter-personal, and intra-personal reliabilities.

Two swimming rubrics (see Appendix D and E) were created for the study. The 10-level rubrics were created by modifying a swimming instruction guideline, Aqua Quest (CRCS, 1996). This work was conducted by the primary researcher with input from the physical education teacher at Mackay Centre School and a teaching assistant. The rubrics were specialized for an inclusive setting; hence swimming with floatation devises was included. According to Baker, O'Neil, and Linn (1993), valid authentic assessment tools have five characteristics:

1. They have meaning for both students and teachers for serving as motivation for

performance.

2. They require demonstration of complex cognition.
3. They exemplify current standards of content quality.
4. They minimize the effects of irrelevant skills.
5. They possess explicit standards for rating or judgment.

The rubrics satisfied the first item, because most of the performance criteria and skill levels of the rubrics were derived from Aqua Quest, which was the curriculum of the swimming classes at the school. Although students did not have the Aqua Quest levels in writing during the class time, they were aware of these levels because the teacher often emphasized individual goals in accordance with Aqua Quest steps. The rubrics were embedded in the curriculum and both teacher and students appeared to enjoy using them.

Although the rubrics were created only for assessment in the current study and not for instruction, it is argued that the rubrics required demonstration of complex cognition such as problem solving, knowledge representation, explanation, applicable to important problem areas (Baker et al., 1993). Each performance criteria of the rubrics was comprised of four major parts: Arms, Feet, Head/Body, and Breathing. Students had to critically judge themselves and peers in using the rubrics.

Most of the test items of the rubrics were derived from the Aqua Quest. Aqua Quest is a typical model of swimming guidelines and this represents current standards of content quality. Although prerequisite skills are not always necessary for subsequent skills for learners with a physical disability (Gelinas & Reid, 2000), it is obvious that Aqua Quest has clear and systematic criteria that are accepted by most swim instructors. Particularly, floatation devices were included in rubric B for children who could not swim independently. This was an addition to Aqua Quest that was believed to be essential by

two experts, the physical education teacher and teaching assistant.

Each level of the rubrics clearly described performance criteria, and it did not require students to have, for example, high level of writing skills to assess performances. Thus, the rubrics required students to simply check the level of performance, therefore minimizing the effects of irrelevant skills (e.g. writing).

In order to enhance accuracy of judgment and to reduce risk of personal bias, the rubrics possessed explicit standards. Both rubrics A and B had 10 levels of performance criteria, and each level had key elements to differentiate it from the prior and subsequent levels. For example, Tuna (8) level of the rubric A (see appendix D) was definitely distinguishable from the prior level, Blowfish (7) because it required swimmers to breathe to side, and it was differentiated from the subsequent level, Dolphin (9) because the Dolphin level required elbows go up when arms pull the water. Also, each key element of the skill was printed in bold letters to distinguish it, which was stressed when teaching students how to use the rubrics.

Reliability

The study was designed to explore how competently students with and without a physical disability could assess performances of peers or themselves. In order to obtain evidence of this, inter- and intra-personal reliabilities were established. Inter-personal reliability can be improved by having detailed rubrics and intense training with the assessors (Joyner & McManis, 1997). As already discussed, both rubrics A and B had detailed criteria, and each level had key elements for accurate judgment. In addition, students had to participate in one or two teaching rubric sessions prior to participation. Eighty percents efficiency was applied as criteria. By demonstrating 80% competence,

it was assumed that if students produced different assessment scores over two weeks or different scores from their teachers, it was not because they did not understand how to use the rubrics. High correlations among assessors, teachers, peers, and self, were obtained to support the objectivity of the rubrics (see Objectivity below).

In contrast to inter-personal reliability, intra-personal reliability is the consistency of measurement over repeated occasions (Joyner & McManis, 1997). In test-retest temporal stability, the three teachers showed competence using the rubrics (.92) as well as student peers (.83) and self assessment (.70). Williams and Rink (2003) investigated competency using observational scoring rubrics with high school physical education teachers. They found that teachers were as reliable observers and scorers 86.89% agreement with monitoring committee members who were composed of professionals and university faculty members. They also mentioned that those teachers showed competence with limited training time (two days, one day, and none). The present findings add to this growing literature and support the use of rubrics by students as well as teachers.

In comparing intra-personal reliability of students with and without a physical disability as well as male and female students, most student groups provided high correlations on peer and self assessments. It appears that students can consistently assess themselves and peers. However, self assessments of students with a disability and that of male students showed low test-retest correlations. The number of participants could be the factor to reduce the reliability. Only four male students and six students with a physical disability were calculated in the self assessment (see Appendix I and J). A limited number of those students might have skewed the data by producing large different scorings (e.g. student 3 and 10, see Appendix I and J).

Objectivity

Assessment results in using rubrics A and B were compared by three assessors: teacher, peers, and self. The results indicated that assessment scores were not significantly different among the three. It appears that different assessors can use the rubrics to accurately assess performance. This provides support for hypothesis one that stated teacher, peers, and self would produce similar assessment results. Thus, students can assess themselves and their peers like teachers. Traditionally, only teacher assigned grades that were considered true score for student performances (Wiggins, 1998). Students believed it was a teacher responsibility. It seems that the rubrics help students to critically judge themselves and understand their strengths and weaknesses. Thus, they can potentially assume some responsibility for their learning process.

Students showed competence by yielding similar assessment scores to the teachers. Rubrics indicate clear guidelines how to accurately assess performance. With enough time for training, students can use rubrics to assess like teachers. Two students who failed to pass with 80% efficiency on the first day achieved the criteria on the second day. In addition, the researcher found that three of the seven students who were disqualified improved their assessing skills after the second training session but did not reach the 80% criteria. Thus, more students might have reached the 80%, if they had more time to practice. Furthermore, if students participated in the creation of rubrics, they would have had much more competence in using and applying the rubrics. Woods and Anderson (2002) reported that students were much more satisfied with rubrics they had devised. Involving students earlier in the process would highlight what they thought was important in an activity unit.

Hypothesis two stated that disability does not affect assessment results. Given

that students could assess like their teachers, did students with a disability do it similarly? In self-assessment comparisons, both students with and without a physical disability demonstrated similar competence to their teachers. Thus, students with a disability could also assess themselves as well as able-bodied peers and teachers. This is evidence to support hypothesis two that regardless of disability, participants yield similar assessment results. It might be that students with a disability have equal-status relationships with able-bodied peers where they reciprocally guide one another (Sherrill et al., 1994).

In peer-assessment comparisons, there was a statistically significant difference between the two groups at the first week of assessment. Students with a disability provided lower scores than students without a disability. There is no obvious explanation for this finding. However, the two groups showed non significant statistical differences for the second week. This likely suggests that the more time raters have in training, the more accurate are their response over time, despite entering the program after attaining at least 80% efficiency. As such, the objectivity of assessment, which involves personal judgmental scoring, can be increased with more training (Linn & Gronlund, 2000).

According to the third hypothesis, gender differences do not affect assessment results. The data indicated that no significant differences between boys and girls. Each sex also showed similar competence in comparison with their teachers in self assessment. Rose, Larkin, and Berger (1997) investigated perceived competence and global self-worth of children who were poorly coordinated and children who were well coordinated. They found that girls had less positive self-perceptions about physical appearance and athletic competence than boys. Gender appears to be influential in such

self-perceptions. It may appear that this study is not consistent with the findings of Rose and her colleagues. However, making accurate judgments of ones physical performance in a specific task (e.g. front crawl) is different from more global self-perception of athletic competence.

Video assessment is widely used in sports games and physical activities because it has the advantage of being an historical recording. Performers and teachers/coaches can discuss the strengths and weaknesses of performers evident on the video. Also, this method has real merit in observing team games so that coaches can look at many players on one screen to see if practiced formations and team work is realized. Nevertheless, video assessment is not always possible in schools because of the extra time required. Also, it is not completely authentic because video images are filtered by lenses which can distort real images, and judgement of performances can change by angles and the number of cameras. Therefore, self assessments immediately after 20m swim were compared to video self assessment one week after the swim. The correlation coefficient of the two different methods was significant and acceptably high enough (.76) that the results of this study based on video assessment can be considered almost equivalent to natural setting.

Rubric systems have many advantages in physical education classes. Lieberman and Houston-Wilson (2002) stated six main advantages to using rubrics: (1) critically evaluate peers' performances (2) being responsible for learning (3) accommodate heterogeneous classes (4) assess the quality and quantity of movement skills (5) chart own progress, and (6) motivated by improving levels of achievement. It is encouraging that the current findings empirically answered the first of these advantages.

Conclusions

Rubrics can be designed with sufficient validity, reliability, and objectivity. More specifically, there was consistency among teachers, peers, and self assessment and few differences between males and females or people with and without a physical disability. It would appear that rubrics can be implemented in school based physical education programs.

Recommendations for future study

1) Future studies should give additional opportunity for participants to practice using rubrics. If rubrics were incorporated into a swimming program on a regular basis as part of instruction and assessment, it would be an interesting research study to investigate the agreement of teachers, peers, and self assessment.
2) The number of participants should be increased.
3) It could be investigated if rubrics enhance academic achievement, since rubrics are known for motivating students in many ways (e.g., improving levels of skills and chart own progress).
4) Rubrics purport to be efficient tools when applied to heterogeneous classes. It would be meaningful to explore how rubrics work in heterogeneously integrated/inclusive classes, even involving severe disability.
5) Current findings support the notion that rubrics may create equal-status relationships in integrated/inclusive settings. Empirical data are demanded to support this inference.

REFERENCES

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*, 1210-1218.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Block, M. E. (2000). *A teacher's guide to including students with disabilities in general physical education* (2 ed.). Baltimore. MD: Paul H. Brookes Publishing.

Block, M. E., Lieberman, L. J., & Connor-Kuntz, F. (1998). Authentic assessment in adapted physical education. *Journal of Physical Education, Recreation & Dance, 69*(3), 48-55.

Bouffard, M. (1993). The perils of averaging data in adapted physical activity research. *Adapted Physical Activity Quarterly, 10*, 371-391.

Bouffard, M. (2003). Foundations of assessment. In R. D. Steadward, G. D. Wheeler & E. J. Watkinson (Eds.), *Adapted physical activity* (pp. 163-187). Edmonton, Alberta: The University of Alberta Press and The Steadward Centre.

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

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston, MA: Pearson Education.

Canadian Red Cross Society (CRCS). (1996). *Canadian red cross water safety service: Water safety instructor manual* St. Louis, MO: Publisher.

Daniel, W. W. (1978). *Applied nonparametric statistics*. Boston, MA: Houghton Mifflin Company.

DeVries, R., Edmiaston, R., Zan, B., & Hildebrandt, C. (2002). What is constructivist

education? Definition and principles of teaching. In R. DeVries, B. Zan, C. Hildebrandt, R. Edmiaston & C. Sales (Eds.), *Developing constructivist early childhood curriculum: Practical principles and activities* (pp. 35-52). New York, NY: Teachers College, Columbia University.

Dewey, J. (1897). My Pedagogic Creed. Published in *The School Journal, 54,* 77-80. retrieved at http://www.infed.org/archives/e-text/e-dew-pc.htm.

Dewey, J. (1934). *Art as experience*. New York, NY: Minton & Balch.

Doolittle, S., & Fay, T. (2002). *Assessment series k-12 physical education: Authentic assessment of physical activity for high school students*. Reston, VA: National Association for Sport and Physical Education.

Field, A. (2000). *Discovering statistics using spss for windows*. London, England: SAGE Publications.

Fosnot, C. T. (1996). Constructivism: A psychological theory of learning. In C. T. Fosnot (Ed.), *Constructivism: Theory, perspectives, and practice* (pp. 8-33). New York, NY: Teachers College, Columbia University.

Gabler, I. C., & Schroeder, M. (2003). *Seven constructivist methods for the secondary classroom: A planning guide for invisible teaching*. Boston, MA: Allyn and Bacon.

Gagnon, G. W., & Collay, M. (2001). *Designing for learning: Six elements in consturctivist classrooms*. Thousand Oaks, CA: Corwin Press.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences* (10th anniversary ed.). New York, NY: Basicbooks.

Gelinas, J. E., & Reid, G. (2000). The developmental validity of traditional learn-to-swim progressions for children with physical disabilities. *Adapted Physical Activity Quarterly, 17*, 269-285.

Gibbons, S. L., & Robinson, B. A. (2005). Student-friendly rubrics for personal and social learning in physical education. *Physical and Health Education, 70*(4), 4-9.

Goodrich, H. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-17.

Gratz, D. B. (2000). High standards for whom? *Phi Delta Kappan, 81*, 681-687.

Hall, E. W., & Salmon, S. J. (2003). Chocolate chip cookies and rubrics: Helping students understand rubrics in inclusive setting. *Teaching Exceptional Children, 35*(4), 8-11.

Hensley, L. D. (1997). Alternative assessment for physical education. *JOPERD, 68*(7), 19-24.

Houston-Wilson, C., Dunn, J. M., Mars, H. v. d., & McCubbin, J. (1997). The effect of peer tutors on motor performance in integrated physical education classes. *Adapted Physical Activity Quarterly, 14*, 298-313.

Jackson, C. W., & Larkin, M. J. (2002). Teaching students to use grading rubrics. *Teaching Exceptional Children, 35*(1), 40-44.

Janesick, V. J. (2001). *The assessment debate: A reference handbook.* Santa Barbara, CA: ABC-CLIO.

Joyner, A. B., & McManis, B. G. (1997). Quality control in alternative assessment. *JOPERD, 68*(7), 38-40.

Katz, J., & Bruning, N. P. (1981). *Swimming for total fitness, a progressive aerobic program.* Garden City, NY: Doubleday & company.

Kirk, M. F. (1997). Using portfolios to enhance student learning & assessment. *JOPERD, 68*(7), 29-33.

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice* (7 ed.). Danvers, MA: John Wiley & Sons.

Lieberman, L. J., Dunn, J. M., Mars, H. v. d., & McCubbin, J. (2000). Peer tutor's effects on activity levels of deaf students in inclusive elementary physical education. *Adapted Physical Activity Quarterly, 17*, 20-39.

Lieberman, L. J., & Houston-Wilson, C. (2002). *Strategies for inclusion - a handbook for physical educators*. Champaign, IL: Human Kinetics.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8 ed.). Upper Saddle River,NJ: Prentice Hall.

Lund, J. (1997). Authentic assessment: Its development & applications. *JOPERD, 68*(7), 25-28, 40.

Lunsford, E., & Melear, C. T. (2004). Using scoring rubrics to evaluate inquiry. *Journal of College Science Teaching*(Sept), 34-38.

Martin, J. E., & Marshall, L. H. (1995). Choicemaker: A comprehensive self-determination transition program. *Intervention in School and Clinic, 30*, 147-156.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5-11.

Montgomery, K. (2000). Classroom rubrics: Systematizing what teachers do naturally. *The Clearing House, 73*, 324-328.

Montgomery, K. (2001). *Authentic assessment, a guide for elementary teachers*. New York, NY: Addison Wesley Longman.

Montgomery, K. (2002). Authentic tasks and rubrics: Going beyond traditional assessments in college teaching. *College Teaching, 50*(1), 34-39.

NASPE. (1995). *Moving into the future: National standards for physical education: A guide to content and assessment*. Renton, VA: Author.

Pitetti, K. H., Jongmans, B., & Fernhall, B. (1999). Feasibility of a treadmill test for

adolescents with multiple disabilities. *Adapted Physical Activity Quarterly, 16*, 362-372.

Popham, W. J. (1997). What's wrong - and what's right - with rubrics. *Educational Leadership, 55*(2), 72-75.

Popham, W. J. (2005). *Classroom assessment: What teachers need to know* (4 ed.). Boston, MA: Allyn and Bacon.

Quebec. Ministry of Education. (2001). Teacher Training: Orientations Professional Competencies. Quebec, QC: National library of Quebec.

Rose, B., Larkin, D., & Berger, B. G. (1997). Coordination and gender influences on the perceived competence of children. *Adapted Physical Activity Quarterly, 14*, 210-221.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment in special and inclusive education* (9 ed.). Boston, MA: Houghton Mifflin Company.

Schultz, R. A. (2002). Teachers as learners: Studying a three-phased rubric assessment plan. *Gifted Child Today, 25*(4), 38-45, 65.

Sherrill, C., Heikinaro-Johansson, P., & Slininger, D. (1994). Equal-status relationships in the gym. *Journal of Physical Education, Recreation & Dance, 65*, 27-31, 56.

Smith, T. K., & Cestaro, N. G. (1998). *Student-centered physical education, strategies for developing middle school fitness and skills*. Champaign, IL: Human Kinetics.

Sova, R. (1995). *Water fitness after 40*. Champaign, IL: Human Kinetics.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50*, 912-927.

Stutzman, R. Y., & Race, K. H. (2004). Emrf: Everyday rubric grading. *Mathematcis Teacher, 97*(1), 34-39.

Thomas, J. R., & Nelson, J. K. (2001). *Research methods in physical activity* (4 ed.). Champaign, IL: Human Kinetics.

Vygotsky, L. S. (1978). *The mind and society: The developmental of higher psychological process*. Cambridge, MA: Harvard University Press.

Ward, A. W., & Murray-Ward, M. (1999). *Assessment in the classroom*. Belmont, CA: Wadsworth Phublishing Company.

Webster, G. E. (1987). Influence of peer tutors upon academic learning time-physical education of mentally handicapped students. *Journal of Teaching in Physical Education, 6*, 393-403.

Wehmeyer, M. L., Sands, D. L., Doll, B., & Palmer, S. (1997). The development of self-determination and implications for educational interventions with students with disabilities. *International Journal of Disability, Development and Education, 44*(4), 305-328.

Welkowitz, J., Ewen, R. B., & Cohen, J. (2002). *Introductory statistics for the behavioral sciences* (5 ed.). New York, NY: John Wiley & Sons.

Whittaker, C. R., Salend, S. J., & Duhaney, D. (2001). Accessing the curriculum: Creating instructional rubrics for inclusive classrooms. *Teaching Exceptional Children, 34*(2), 8-13.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*, 200-214.

Wiggins, G. (1998). *Educative assessment, designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Williams, L., & Rink, J. (2003). Teacher competency using observational scoring rubrics. *Journal of Teaching in Physical Education, 22*, 552-572.

Woods, M. J., & Anderson, D. (2002). Students designing and applying evaluation rubrics in an aerobics unit. *The Physical Educator, 59*(1), 38-56.

Yun, J., & Ulrich, D. A. (2002). Estimating measurement validity: A tutorial. *Adapted Physical Activity Quarterly, 19*, 32-47.

Appendix B

Parent Consent Form

Hello

I am a graduate student in Kinesiology and Physical Education at McGill University. As a requirement for achieving a Master's Degree I am conducting a research project at the Mackay Center School. I have been working at the MacKay Center School in the pool and gym for the past 3 years as a physical education assistant to Bob Simpson. Consequently, I have come to know the majority of the students who attend the school, including your child. McGill requires a letter of consent whenever research is conducted involving students, stating the purpose, procedures, and conditions of the research. This does not imply that the project involves any risk; the intention is simply to assure the respect and confidentiality of the individuals concerned.

The study is designed to see if students can assess the performance of their peers as well as themselves with a rubric (a sample rubric is attached on the other side of this letter). Participation in the study requires your child to join in one or two learning sessions to understand how to use rubrics during lunch hour. After the learning sessions, your child will be asked to perform 20-meter (two-width of Mackay pool) front crawl swimming which will be videotaped, then asked to return to a room on the following week during the lunch hour to watch the videotape and assess the swimming performances with the rubric. Swimming takes place during the part of the regular swimming class, and the assessment will be held once a week for two weeks.

Your child's participation in this study is strictly voluntary. You or your child may refuse to participate or discontinue participation at any time without explanation. If you decide not to participate, or if you discontinue your participation, you will suffer no prejudice regarding educational services.

All information obtained during this study will be kept strictly confidential. Your child will be identified by an ID number and the swimming performance videotapes and any assessment documents will be locked in a filing cabinet in the investigator's office. The results from this study may be published; however, the identity of your child will not be revealed since only groups will be designated. In order to verify the research study data, monitors from the Research Ethics Board at McGill may review the research files in order to verify compliance with institutional research regulations.

By signing on next page, you are indicating consent for your child to participate in study, including videotaping. As well, you are confirming that you have read the above information and that you are aware of the nature and demands of the study.

I want to participate in the study.

Student: _____ Date: _____

Please return the signed form to Bob Simpson.

Appendix D

Scoring Rubric A (20m front crawl)

Assessor:_____ **Performer:**_____ Date:____/ May / 2006____

	CRITERIA				CHECK
Level	**Arms**	**Feet**	**Head/Body**	**Breathing**	
Starfish	--	**▪Walks**	--	--	
Sea Urchin	--	**▪Push / glide**	--	--	
Jellyfish	**▪Pull under body**	**▪Legs moving**	**▪Face in water**	--	
Shrimp	**▪Reach in front of head**	**▪Kick** in water	▪Face in water	--	
Seahorse	▪Reach in front of head **one at a time**	▪Kick with splash	▪Face in water	--	
Goldfish	▪Reach in front of head one at a time	▪Kick with splash	▪Face in water	**▪Head up breathe** **▪Blow bubbles in water**	
Blowfish	▪Reach in front of head one at a time **above water**	▪Kick with splash	▪Face in water	▪Head up breathe ▪Blow bubbles in water	
Tuna	▪Reach in front of head one at a time above water	▪Kick with splash **▪Toes pointed**	**▪Body straight** **▪Head straight**	**▪Breathe to side** ▪Blow bubbles in water	
Dolphin	▪Reach in front of head **▪Elbows up** **▪Pull past hips**	▪Kick with splash ▪Toes pointed	▪Body straight ▪Head straight **▪No hips sway**	▪Breathe to side ▪Blow bubbles in water	
Shark	▪Reach in front of head one at a time above water ▪Pull past hips **▪Pull in S-pattern**	▪Kick with splash ▪Toes pointed	▪Body straight ▪Head straight ▪No hips sway	▪Breathe to side ▪Blow bubbles in water	

Appendix E

Scoring Rubric B (20m front crawl)

Assessor:_____ **Performer:**_____ Date:____/ May / 2006__

Level	CRITERIA					CHECK
	Devices	Arms	Feet	Head/Body	Breathing	
Starfish	**Caregiver**	--	--	--	--	
Sea Urchin	**Life Jacket & Noodle**	--	--	**Float**	--	
Jellyfish	Life Jacket & Noodle	**▪Hold the life jacket**	**▪Legs moving**	--	--	
Shrimp	**Life Jacket**	**▪Pull under body**	▪Legs moving	**▪Face in water**	--	
Seahorse	Life Jacket	**▪Reach in front of head one at a time**	**▪Kick with splash**	▪Face in water	--	
Goldfish	**Noodle**	▪Reach in front of head one at a time	▪Kick with splash	▪Face in water	**▪Head up breathe** **▪Blow bubbles in water**	
Blowfish	Noodle	▪Reach in front of head one at a time **above water**	▪Kick with splash	▪Face in water	▪Head up breathe ▪Blow bubbles in water	
Tuna	**Belt**	▪Reach in front of head one at a time above water	▪Kick with splash **▪Toes pointed**	**▪Body straight** **▪Head straight**	▪Head up breathe ▪Blow bubbles in water	
Dolphin	Belt	▪Reach in front of head **▪Elbows up** **▪Pull past hips**	▪Kick with splash ▪Toes pointed	▪Body straight ▪Head straight **▪No hips sway**	**▪Breathe to side** ▪Blow bubbles in water	
Shark	Belt	▪Reach in front of head ▪Elbows up ▪Pull past hips **▪Pull in S-pattern**	▪Kick with splash ▪Toes pointed	▪Body straight ▪Head straight ▪No hips sway	▪Breathe to side ▪Blow bubbles in water	

Appendix F

Raw data of the 1st week: Video assessment

Assesors	Performers																			
	S1	**S2**	**S3**	**S4**	**S5**	***** A1**	**S6**	**S7**	**S8**	**S9**	**S10**	**S12**	**A2**	**A3**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**A4**
* T1	10	9	8	8	4	3	10	9	8	8	8	6	9	8	8	8	9	10	8	7	6
T2	9	8	8	6	4	3	9	9	8	8	8	3	8	5	8	8	8	9	8	7	5
T3	9	9	7	6	4	4	9	9	9	9	8	4	9	6	9	8	9	10	8	7	5
** S1	**8**	7	7	6	4	3															
S2	8	**10**	8	7	5	6															
S3	2	4	**5**	3	4	5															
S4	8	7	6	**5**	5	4															
S5	9	7	3	4	**4**	2															
S6							**8**	8	7	8	7	3	6	8							
S7							8	**8**	7	8	6	6	6	7							
S8							8	6	**7**	8	6	5	9	8							
S9							8	9	7	**8**	8	6	7	7							
S10							5	5	6	7	**6**	5	5	5							
**** *S11*							8	9	5	4	7	6	10	8							
S12							8	8	8	9	9	**7**	8	6							
S13															**8**	7	8	9	8	7	6
S14															8	**8**	8	10	7	5	4
S15															8	8	**9**	9	8	8	4
S16															9	7	8	**10**	10	7	5
S17															8	8	8	9	**9**	8	4
S18															8	8	6	8	9	**5**	6
S19															7	7	8	9	7	7	3

* T, Teacher 1, 2, and 3

** S, Student 1-19

*** A, Additional student 1-4 only for swimming

**** *Italic (S11 and S19)* - Students did not participated in swimming

Appendix G

Raw data of the 2nd week: Video assessment

Assesors	Performers																			
	S1	**S2**	**S3**	**S4**	**S5**	**A1**	**S6**	**S7**	**S8**	**S9**	**S10**	**S12**	**A2**	**A3**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**A4**
T1	9	9	8	8	4	3	10	9	8	8	8	7	9	8	9	8	9	10	8	7	6
T2	9	8	8	6	4	3	9	9	8	8	8	3	8	8	8	8	8	9	8	7	5
T3	9	9	8	7	4	4	9	9	8	8	7	4	8	7	9	8	9	10	8	7	5
S1	**8**	7	5	6	4	3															
S2	8	**10**	7	8	5	5															
S3	10	8	**10**	10	9	10															
S4	8	7	7	**7**	5	5															
S5	9	10	6	7	**4**	3															
S6							**8**	7	7	8	8	7	7	7							
S7							8	**7**	7	8	7	7	7	7							
S8							6	5	**6**	8	6	4	9	8							
S9							8	7	7	**8**	8	7	8	7							
S10							4	5	5	4	**3**	4	3	4							
S11							8	6	4	9	5	4	8	8							
S12							8	8	8	8	9	**6**	10	6							
S13															**8**	8	8	9	8	7	4
S14															7	**6**	9	10	9	5	6
S15															6	8	**10**	10	9	6	4
S16															9	7	9	**10**	8	6	4
S17															8	7	9	9	**9**	8	4
S18															8	8	8	8	9	**5**	5
S19															8	7	7	9	8	6	4

* T, Teacher 1, 2, and 3

** S, Student 1-19

*** A, Additional student 1-4 only for swimming

**** *Italic (S11 and S19)* - Students did not participated in swimming

Appendix H

Median data of Teachers, Peers, and Self assessments

Assessors	Teachers (N=3)		Peers (N=19)		Self (N=17)		
Performers	*** W1	W2	W1	W2	W1	W2	**** NO video
* S1	9.0	9.0	8.0	8.5	8.0	8.0	8.0
S2	9.0	9.0	7.0	7.5	10.0	10.0	10.0
S3	8.0	8.0	6.5	6.5	5.0	10.0	8.0
S4	6.0	7.0	5.0	7.5	5.0	7.0	5.0
S5	4.0	4.0	4.5	5.0	4.0	4.0	4.0
** A1	3.0	3.0	4.0	5.0			
S6	9.0	9.0	8.0	8.0	8.0	8.0	8.0
S7	9.0	9.0	8.0	6.5	8.0	7.0	9.0
S8	8.0	8.0	7.0	7.0	7.0	6.0	7.0
S9	8.0	8.0	8.0	8.0	8.0	8.0	8.0
S10	8.0	8.0	7.0	7.5	6.0	3.0	8.0
S12	4.0	4.0	5.5	5.5	7.0	6.0	5.0
A2	9.0	8.0	7.0	8.0			
A3	6.0	8.0	7.0	7.0			
S13	8.0	9.0	8.0	8.0	8.0	8.0	9.0
S14	8.0	8.0	7.5	7.5	8.0	6.0	8.0
S15	9.0	9.0	8.0	8.5	9.0	10.0	8.0
S16	10.0	10.0	9.0	9.0	10.0	10.0	9.0
S17	8.0	8.0	8.0	8.5	9.0	9.0	8.0
S18	7.0	7.0	7.0	6.0	5.0	5.0	6.0
A4	5.0	5.0	4.0	4.0			
Mean	7.38	7.52	6.86	7.10	7.35	7.35	7.53
SD	1.96	1.91	1.44	1.35	1.80	2.15	1.62

* S, Student 1-18

** A, Additional student 1-4 only for swimming

*** W, Weeks 1 and 2

**** NO video, Scores of immediately after each swimming

Appendix I

Median data of peer and self assessments for students with and without disability

Assessors	Peer Assessment				Self Assessment			
	*** SWD (N=7)		**** SWOD(N=12)		Self WD (N=6)		Self WOD (N=11)	
Performers	***** W1	W2	W1	W2	W1	W2	W1	W2
* S1	8.0	9.0	8.0	8.0			8.0	8.0
S2	7.0	8.0	8.5	8.5			10.0	10.0
S3	5.0	7.0	7.5	6.0	5.0	10.0		
S4	4.0	7.0	6.5	7.0	5.0	7.0		
S5	4.0	5.0	4.5	4.5	4.0	4.0		
** A1	4.0	5.0	4.5	4.0				
S6	6.5	6.0	8.0	8.0			8.0	8.0
S7	6.5	6.5	8.0	7.0			8.0	7.0
S8	7.0	6.5	7.0	7.0			7.0	6.0
S9	8.0	6.0	8.0	8.0			8.0	8.0
S10	7.5	6.0	7.0	7.0	6.0	3.0		
S12	6.0	5.0	6.0	7.0	7.0	6.0		
A2	6.5	6.5	7.0	8.0				
A3	5.5	5.0	8.0	7.0				
S13	7.5	8.0	8.0	8.0			8.0	8.0
S14	7.5	7.5	8.0	7.0			8.0	6.0
S15	7.0	7.5	8.0	9.0			9.0	10.0
S16	8.5	8.5	9.0	10.0			10.0	10.0
S17	8.0	8.5	8.0	9.0			9.0	9.0
S18	6.0	5.5	7.0	6.0	5.0	5.0		
A4	4.5	4.5	4.0	4.0				
Mean	6.40	6.60	7.17	7.14	5.33	5.83	8.45	8.18
SD	1.43	1.34	1.37	1.58	1.03	2.48	0.93	1.47

* S, Student 1-18

** A, Additional student 1-4 only for swimming

*** SWD, Students with a disability

**** SWOD, Students with out a disability

***** W, Weeks 1 and 2

Appendix J

Median data of peer and self assessments for male and female students

Assessors	Peer Assessment				Self Assessment			
	Male (N=5)		Female (N=14)		Male (N=4)		Female (N=13)	
Performers	*** W1	W2	W1	W2	W1	W2	W1	W2
* S1	8.0	8.0	8.0	9.0			8.0	8.0
S2	8.5	8.5	7.0	8.0	10.0	10.0		
S3	7.0	7.0	5.0	6.0			5.0	10.0
S4	6.0	7.5	4.0	7.0	5.0	7.0		
S5	5.0	5.0	4.0	4.0			4.0	4.0
** A1	5.0	5.0	3.0	3.0				
S6	5.0	4.0	8.0	8.0			8.0	8.0
S7	5.0	5.0	8.0	7.0			8.0	7.0
S8	6.0	5.0	7.0	7.0			7.0	6.0
S9	7.0	4.0	8.0	8.0			8.0	8.0
S10	6.0	3.0	7.0	7.5	6.0	3.0		
S12	5.0	4.0	6.0	6.5			7.0	6.0
A2	5.0	3.0	7.5	8.0				
A3	5.0	4.0	7.5	7.0				
S13	7.5	8.0	8.0	8.0			8.0	8.0
S14	7.5	7.5	8.0	7.0			8.0	6.0
S15	7.0	7.5	8.0	9.0			9.0	10.0
S16	8.5	8.5	9.0	10.0			10.0	10.0
S17	8.0	8.5	8.0	9.0			9.0	9.0
S18	6.0	5.5	7.0	6.0	5.0	5.0		
A4	4.5	4.5	4.0	4.0				
Mean	6.31	5.86	6.76	7.10	6.50	6.25	7.62	7.69
SD	1.33	1.93	1.72	1.77	2.38	2.99	1.61	1.84

* S, Student 1-18

** A, Additional student 1-4 only for swimming

*** W, Weeks 1 and 2