A Study of the Reliability and Validity of the Standardized Exams Used in Grade 10 Science in a Winnipeg School Division

by

Jason H. Braun

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF EDUCATION

Department of Educational Administration, Foundations and Psychology

University of Manitoba

Winnipeg, Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

A Study of the Reliability and Validity of the Standardized Exams Used in Grade 10 Science in a Winnipeg School Division

BY

Jason H. Braun

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

Master of Education

ABSTRACT

The purpose of this study was to examine the reliability and validity of Grade 10 science exams administered in 2006 and 2007 in a school division in the province of Manitoba. Over 1300 hundred grade 10 science students from five high schools participated in the study. The data was gathered from a secondary analysis of exam results using ExamSystem II software.

A review of the examination policies in the division suggest the purpose of the grade 10 science exams did not change during the 2006 and 2007 testing years. In addition, a number of statistical analyses of the items in the four examinations showed that they were very similar in all four testing periods. The statistical analyses of the exams suggest that the change in the exam development process from 2006 to 2007 did not affect the quality of the exam or the achievement of the students. Finally, the analyses of the students' test scores suggest that the type of invigilation practice did not affect student achievement.

In all respects, the standardized exams in the school division are viewed by the teachers, administrators, students, parents, and trustees as being an effective accountability instrument that provides reliable and valid information. There is, in addition, considerable evidence that these exams are important in developing divisional policies about the evaluation of students, and it is recommended that other divisions consider developing similar policies and procedures for assessing their students.

ACKNOWLEDGEMENTS

"To boldly go where no one has gone before". My trek to complete this study could not have happened without a solid crew surrounding me with support. Firstly, Janine, my loving partner and colleague. Thank you for knowing when to back off and when to force a course correction. To my children, Jude, Jordan, and Jersey, who kept me grounded and remind me to the important things in life. I hope my study will compensate for my occasional absences by having each of you reach your maximum potential in the schools you attend.

To Dr. Rodney Clifton – my advisor, whom I cannot begin to thank for all the time, energy, and caring that he devoted to my project and to my education. Thank you for being patient and also demanding when needed. You were an excellent navigator in my trek.

To Dr. John Long – a fantastic role mode as professor, colleague, and friend. The university will sorely miss your expertise and wealth of knowledge.

To Dr. Karen Smith – thank you for your guidance with the manuscript and challenging the social context of the study.

To the St. James-Assiniboia School Division – thank you for initiating a Master's Cohort and encouraging its teachers to pursue their studies in Education. Without such positive support, my trek may never have happened.

To my brother Curtis–thank you for the education debates around the coffee table. We may not have always agreed but the "discussions" were invaluable.

Lastly, to my mother, Donna Jaycee Braun – although you were not with me during this trek, I feel your spirit was always there. Such is, a fair and foul relationship.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In Macbeth, Shakespeare wrote "fair is foul and foul is fair" (Act I, Scene I), which reflects today's opposing views on standardized examinations. Academics opposed to standardized exams, such as Alfie Kohn, suggest that standardized exams are not a fair assessment instrument, but a foul stick to beat instruction and learning. Academics in favour of standardized testing, such as Richard Phelps, suggest that standardized exams have fair consequences in assessment, and are not a foul stick to beat instruction and learning. This thesis is an assessment of the fairness of the Grade 10 science standardized examinations developed and used to measure the degree of mastery by students of the science curriculum in the Galileo School Division in Winnipeg, Manitoba, Canada.

Grade 10 students are required by provincial policy to take and pass the Grade 10 science course as part of requirements to receive a provincial high school diploma (Manitoba Education, Citizenship & Youth, 2007a). The "SC20F" abbreviation will be used to represent the course. The "SC" refers to science and "20F" indicates that the course is at the grade ten level. Since 1999, the Galileo School Division requires that all students enrolled in SC20F to write a standardized examination on the content of the course at the end of the semester worth 25% of their final grades. These exams, as well as all the others, are under the administration of the division's Assessment and Evaluation Coordinator who reports to the Assistant Superintendent of Programming and Curriculum. In developing the exams, the Assessment and Evaluation Coordinator sends a set of guidelines to the curriculum coordinator, based on divisional policy concerning assessment and evaluation. The Science Coordinator selects a science teacher to construct the SC20F exam. The science teacher

selection process begins with a "call for setters" notice sent out to every science teacher currently teaching grade 10 science. A number of qualifications are considered when selecting the teacher to set the exam. The teacher should have at least two years experience teaching the course with some experience in reviewing divisional science exams in previous years. The science teacher should also demonstrate the ability to adhere to deadlines for submissions. Although it may appear to be easier to allow the same science teacher to set the exam from year to year, it is important to select a number of teachers for their professional development and for the fairness of the divisional exam development process. The science teacher assigned to set the exam uses an exam template and divisional established guidelines as a framework. The exam template (see Appendix C) itself is developed by a committee of science teachers under the guidance of the Science Coordinator, and the divisional guidelines (see Appendix A) are developed by the Assessment and Evaluation Coordinator in accordance with the evaluation policies. Following the development of the exam, it is reviewed by two other science teachers for curriculum correspondence, word usage, and grammar, before being submitted to the Assessment and Evaluation Coordinator for a final review and printing.

The SC20F exam is administered twice during each school year, once in January and once in June, and both exams are constructed by the same science teacher. It is usual that in constructing the June exam, after the January exam is written by the students, the test items are refined by the exam setter based on the teachers' written feedback from the previous exam. This particular way of exam construction along with the science teachers' limited background in psychometrics, raises questions of the reliability and validity of the science exams.

The Purpose of the Study

The purpose of the study was to assess the fairness and utility of the standardized testing program used in Grade 10 Science as a summative assessment in the Galileo School Division in Winnipeg, Manitoba. Summative assessment, or "assessment of learning", measures how well students have learned the material in the curriculum at a particular point in time (MECY, 2008). Assessments of learning are implemented after learning is supposed to have occurred to determine if it did, in fact, occur. The results of these assessments are used to make statements about the students' learning status at a particular point in time, and to inform the students, their parents, their teachers and other people outside the school and division about the students' level of achievement in the course. Of course, divisional exams must be both reliable and valid in order to be credible and useful to all those interested in the achievement of the students. Manitoba Education, Citizenship, and Youth (2006) states that exams must also:

1. Imbed curriculum learning outcomes in the structure and design of evaluation tools;
2. Report student achievement and progress;
3. Inform students and others about progress;
4. Use pre-determined standards or levels of achievement; and
5. Involve parents in discussing learning growth and support; i.e., strengths, areas needing improvements, and goals.

An important characteristic of any assessment instrument is its quality. In line with this, Phelps (2007, p. 89) notes that often an assessment is hastily constructed by teachers and

never revised once it has been administered. Nevertheless, good assessment, as noted above, requires a high degree of validity and reliability. Validity is the "extent to which a test measures the quality it purports to measure" and reliability refers to "the degree to which test scores are consistent across time, conditions, and test-takers" (Phelps, 2007, p. 88). To insure that a test has a high degree of validity and reliability, the test development process must include:

1. A well defined purpose;
2. Carefully developed test items; and
3. Consistent test administration and scoring procedures.

The Purpose of Tests

The purpose of tests should be clearly defined by the agency developing the instrument, and should describe the intended goal of the test, the population of students to be tested, and a description of what will be done with the results of the test (McMillan, 2007). A criterion-referenced standardized test measures students' achievement against a set of predetermined standards established by committees of teachers and experts in a discipline, as noted above. In addition, standardized test items are pre-tested to identify and remove possible ambiguities and they would be marked by specially trained teachers following established protocols to insure that the grading is consistent for every student. Consequently, a criterion-referenced standardized test is "a relatively objective test that yields the same score for all students who achieve the same performance outcome, irrespective of their school, school board, or province" (Cirtwell, Clifton, & D'Orsay, 2002, p. 16). Phelps (2007, p. 4) reinforces this point when he notes that a criterion-reference standardized test

measures achievement with respect to a set of specific standards, that is "an achievement test designed to cover a specified content domain that is usually identified by content standards." Similarly, Traub (1994, p. 20) says that a standardized test "is designed to assess the knowledge and understanding a student has acquired of a school subject." Closer to the concerns of citizens in this province, Manitoba Education, Citizenship, and Youth (2007b, p. 1) informs teachers that:

> Standards [standardized] tests are intended to provide pertinent information about each student's knowledge and skills in relation to student learning outcomes as set out in provincial curriculum documents. To provide students with an accurate, balanced, and well-rounded profile of their progress and achievement, standards tests complement a variety of classroom assessments. These assessments may include teacher observations, writing samples, exhibitions, portfolio assessments, and demonstrations.

In a similar vein, the Galileo School Division (2000) testing policy has the following goals:

1. To provide an additional assessment component, on a wider basis than traditional classroom-based assessments, for individual student evaluation;
2. To provide feedback to staff for reference in the continuous improvement of curricular and instructional practices;
3. To provide information to the general public;
4. To ensure adherence to prescribed curricula; and
5. To assist the school division in improving the quality of academic standards.

In essence, this divisional policy states that the standardized testing is used to measure the achievement of students, and that the results are used to provide feedback to the student,

parents, teachers, divisional administrators, and the general public. The feedback is in the form of a report card to the students and their parents and school profile reports to administrators and citizens. The policy also states that the standardized testing is used to ensure that the teachers follow the provincial curriculum, and that instructional practices are adjusted as a consequence of the data provided by the testing. Even though the grade 10 science standardized examination is written by all grade 10 students in the division, there are, of course, exceptions in the case of adaptive or enriched programming for about 25 students representing approximately 8% of the 400 students who write the test every semester.

Test Development

According to Phelps (2007), criterion-referenced test items "must be clear, based on subject matter the students have been taught, and be fully answerable within the time frame allotted for the response." The test item development for the SC20F examination follows these guidelines as well as those established by the Assessment and Evalution Coordinator. Appendix A contains the guidelines for the 2007 Divisional tests. These guidelines, of course, are based on division-wide policies on standardized testing, including the prescribed weighting and format of the exam. Each standardized exam administered by the Division is differentially weighted depending on the grade level and the subject, and according to that policy each exam must have both selected response and free response items. The school division has had standardized science exams in grades 8, 10, 11, and 12 since 1990. The grade 8 science exam is worth 20% of the students' final grade and has a writing time of two hours. The exam is administered at the end of the school year in June. In grade 10, the science students write a standardized exam worth 25% of their final grades, with a writing

time of 2.5 hours, and this exam is administered at the end of each semester. For grade 11 biology, chemistry, and physics, the students have standardized exams worth 25% of their final grades, the writing time in three hours, and these exams are administered at the end of each semester. Finally, biology, chemistry, and physics courses in grade 12 have standardized exams worth 30% of the students' final grades, with a writing time of three hours, and they are written at the end of each semester. The percentage weighting of the examinations has been established by the Board of Trustees and has not been adjusted since 1997.

The percentage weightings of the selected response items and free response items are also established by the Board. In recent years, the Assessment and Evaluation Coordinator has allowed the percentage weightings to vary for the optional science courses so that more weight is placed on the free response items than on the multiple-choice items. The change was made to reflect the theory of authentic assessment because selected response items were not seen by the Trustees, at least, as being authentic enough in comparison with the free response items. Table 1 presents the descriptive characteristics of the SC20F examinations written in January and June of the 2006 and 2007 test year.

Table 1

Characteristics of the SC20F Exams in January 2006 and June 2007

Exam date	Number of selected response items	Weighting of selected response items	Number of free response items	Weighting of free response items
January 2006	100 items	65%	14 items	35%
June 2006	100 items	65%	15 items	35%
January 2007	75 items	65%	12 items	35%
June 2007	75 items	65%	12 items	35%

The grade 10 science course in the Galileo School Division follows the provincial science curriculum (SC20F). The Galileo School Division courses are semestered at the high school level with the first semester concluding in January and the second semester finishing at the end of June. The SC20F course is offered in both semesters with standardized examinations at the end of each semester. The grade 10 science test is weighted at 20% of the students' final grade, and is composed of selected-response and constructed-response items, worth 65 percent and 35 percent respectively (see Appendix B for a copy of January 2006 SC20F Exam).

The Assessment and Evaluation Coordinator guidelines for constructing the examination contain two major sections, the first addresses the responsibilities of the exam setter and the second addresses the responsibilities of the two teachers who review the exam. The exam setter constructs the test items and the answer key and the items must be based on the curriculum and provide a range of difficulty from knowledge-based understanding to the

synthesis level of thinking that are reflected in Bloom's taxonomy (Bloom et al., 1956). The exam setter must consider the number of items in the exam, the item analysis of previous exams, and comments made by teachers about the previous exams. The guidelines allow an exam to be developed by a group of three teachers, but in practice, to date, divisional exams have always been set by an individual teacher and reviewed and corrected by two other teachers.

A group-developed exam inherently has a number of logistical problems since this approach requires dividing up the workload, and getting the teachers together on a regular basis. In June the Science Coordinator assigns the setter and the two reviewers based on their interests and their familiarity with the provincial curriculum. Fortunately, the setter and the two reviewers are paid extra for their work. To assist the exam setter in developing the test items, the Science Coordinator chairs meetings of groups of science teachers to help identify the essential learning outcomes in the science curriculum. The meetings involve representatives of about 2 teachers from each of the five high schools, for a total of about 10 teachers from about 13 teachers who teach the course. All test items are reviewed by two teachers and are clarified in terms of depth and breadth. Appendix C contains the template of essential learning outcomes developed during a 2006 SC20F curriculum meeting. Based on the Division's experience in setting these exams and the comments received from other setters, the job of setting an exam is not only very challenging but also time consuming.

The two teachers who are the exam reviewers follow the guidelines provided by the Assessment and Evaluation Coordinator and the template of essential learning outcomes. Their tasks include checking:

1. The accuracy of the content;
2. The accuracy of the answer key;
3. The clarity of the exam wording;
4. Typing errors;
5. That the selected response items are positive; and
6. The length of the examination.

The examination setter is given three months, two of which are July and August, to write the exam items, and then the reviewers are given a week to complete their reviews. The reviewers and the Science Coordinator meet and discuss any problems that have been identified. The Science Coordinator communicates the concerns to the exam setter at a meeting and the setter has the option to reject the suggestions if the setter cannot be convinced why suggested change(s) should be made. The review process is followed, first, with one reviewer and then with the second reviewer. Nevertheless, the final decision on the changes ultimately rests with the Science Coordinator. This particular responsibility has proved difficult to some extent because neither the exams nor the changes satisfy all the science teachers. Nevertheless, the examination is ultimately submitted to the Assessment and Evaluation Coordinator for final review before it is sent to be printed. Unfortunately, very few of the teachers who set or reviewed the SC20F examinations over the last few years have had a background in psychometrics, therefore, as identified in the Assessment and Evaluation Coordinator's guidelines, the item analysis of past exams and the essential learning outcomes template are extremely important in helping the teachers develop good test items.

Test Administration and Scoring

The examinations are administered locally by teachers in each of the high schools, but the time and date of the examination are set by the Assessment and Evaluation Coordinator in consultation with the subject coordinators in each of the five schools. The school administrators are responsible for informing the students and their parents of the examination schedule. This is done through the school's website and newsletter, while the school division also publishes the examination schedule on its' website. Each school principal determines the location in the building where the examination will be written and oversees the supervision of the students writing the exams. The completed exams are given to each subject area teacher, and each of them scores the free response items, afterward reporting the marks to the school administrators and the school division. The selected response bubble sheets are sent to the school division office to be scored by computer and the scores are collated with the scores from the free-response items for reporting to the students and their parents, and for data analyses so that the results can be released to the public. Students may be informed by the subject area teacher of their marks provided there are no outstanding library books or student fees. Otherwise, the exam mark is reported on each student's report card. Data on exam results is compiled and shared with schools in the fall of the next school year by the Division.

The Research Questions

According to Phelps (2007), a valid examination evaluation process will help answer three important questions:

1. Is there evidence of validity that guarantees the test scores are meaningful?
2. Is measurement error minimal to assure reliability? and
3. Are the grades derived from the test fair to all students?

This test quality checklist forms part of the framework of this study of the grade ten criterion-referenced science standardized examination in the Galileo School Division. Specifically, I will examine three main questions:

1. Are the SC20F examinations in January and June, in the same testing year, comparable in purpose and results?

The development process of the SC20F exams changed from the 2006 year to the 2007 year. In 2006, the test development process involved a viewing session of the exam for all interested science teachers, which always included the majority of teachers. The viewing session provided an opportunity for the teachers to review the exam in draft form and to comment on it. The viewing session was held in the Professional Staff Development Centre under the supervision of the Science Coordinator. Copies of the draft exam and the answer key were distributed to the teachers who were then allowed to sit at tables and read through the exam items. Teachers were allowed to discuss items with each other, but they were not allowed to make notes that could be taken out of the room. The draft exams and the keys were collected after the viewing session and they were shredded. Any problems identified during the session were reported to the Science Coordinator for consideration. The purpose of the viewing session was twofold: it allowed teachers to check to see if they covered the

science curriculum in their teaching, and it helped teachers identify problems with the examination, such as language use and the correspondence of the items with the provincial curriculum. These reviewing sessions were valuable because they often identified problems that were missed by the exam reviewers and because the conversations between the teachers and science coordinator improved the understanding of each others' needs and interests with respect to use science curriculum and science education.

Nevertheless, in the 2007 test year, the viewing session was discontinued, apparently for security reasons. The absence of the viewing session placed greater responsibility on the teachers to follow the essential learning outcomes template developed by the division. Thus, this template needed to state very clearly the outcomes that were to be measured, the weighting of the topics in the exam, for example, that the physics portion is worth 25% of the total exam points, and the format of the exam (e.g., the number of selected response compared to free response items). As stated earlier, the development of the exam template was created in a meeting of the teachers from each of the five high schools, chaired by the Science Coordinator. The exam template meetings, however, do not occur on a yearly basis, but are convened when circumstances warrant, such as when new provincial curricula are approved or when exam templates are changed. Notably, the present SC20F exam template was developed in 2005 and has not changed since that date. As a result, the discontinuing of the exam viewing session has created significant concerns among the science teachers. The teachers, in fact, think that the exam setter and the exam reviewers have an advantage because they have seen the exam and, in all likelihood, their students will be more successful because they are better able to prepare for them, even though they are forbidden, of course, from using the exam items in their instruction. The change in the test development process,

that is, the removal of the exam viewing session from the 2006 test year to the 2007 test year, led to the second question for this study:

2. Did the change in the test development process from the 2006 test year to the 2007 test year affect the quality of the exam and the achievement of students?

The Assessment and Evaluation Coordinator is responsible for packaging and delivering the exams to the schools. Contained in each package is an affidavit (see Appendix D) describing the responsibilities of the school administrators when they invigilate the exam. The administrators' signatures on the affidavit indicate that the exam packages containing the exams have not been opened until the date the exam was to be written. The affidavit does not describe the procedures that are to be followed when the exam is written, such as the type of room to be used (e.g. gymnasium or classroom) or if the students are to write at independent desks or large tables. Also, the affidavit does not indicate the number of supervisors or the nature of the supervision (e.g. administrators, department heads, or teachers who taught the course).

The Assessment and Evaluation Coordinator, in consultation with subject coordinators and school administrators, schedules the date and time for the exam, which universally applies to all schools. This information is sent out to school administrators in the form of a calendar (See Appendix E for January 2008 Exam Schedule) that includes the divisional and provincial exam dates and times. The school administrators are responsible for distributing this information to their teachers and students, and for organizing the invigilation of the examinations. There are no guidelines restricting the nature of the invigilation, only that the exam must be administered on the date and time prescribed by the Assessment and Evaluation Coordinator. However, the administration of the exams varies

across the five high schools. Some schools have their students write the exam in a large gymnasium and others have them write the exam in classrooms; some schools have a seating plan and others have "scramble" seating. The variability in the administration of the science exams, a third question is addressed in this thesis:

3. Does the difference in administration and supervision of the exam have an effect on the exam results?

Essentially, this study examines two years of implementation of standardized exams in grade 10 science. Surprisingly, the Galileo School Division is the only school division in the Province of Manitoba that uses standardized exams in the sciences. As such, the study contributes to an understanding of the value of standardized exams and the benefits of having standardized assessment practices. In addition, the study shows the effectiveness of the current practices in the development and use of standardized exams. In this respect, Hirsch (1996, p. 214) states: "Without effective monitoring and high incentives, including high-stakes testing programs, no educational system has achieved or could achieve excellence and equity." Hirsch's comment is also reflected in the work of John Bishop (1992, 1994) in studies associated with standardized exams. His arguments, in fact, illustrate the importance of having standardized measuring instruments which contributes to effective teaching and better learning for students. The three questions posed in this study will help determine the validity and reliability of the SC20F standardized exam, and address the importance of having a standardized invigilation practice when using standardized exams.

Limitations of the Study

Even though, this study is significant in many respects, it obviously has some limitations. There are, in fact, three major limitations to this study. The first limitation is the study is restricted to one school division, Galileo School Division, located in the western part of Winnipeg. The student population of the high schools ranges from 350 to 850 students with approximately 3000 high school students in total. All five high schools in the division were included in this study, but the Galileo School Division is the only school division in the Province that uses standardized exams in grade 10 science. If other school divisions used standardized exams, the comparison of the exam development models, the implementation practices across the province, and the results, would make this study much more valuable. That is, the study could be viewed in a broader context for comparison purposes.

The second limitation is that the study is restricted to the SC20F standardized exams invigilated during the 2006 and 2007 school years. In each of these years, the SC20F exam was written twice, with one version written in January and the second version written in June. Therefore, the data collected for this study includes four SC20F standardized exams from the five high schools, written during the 2006 and the 2007 testing years. Although the Galileo School Division has standardized exams in other science courses, such as biology and chemistry, and in other subjects, such as, English, Mathematics, and Social Studies, these exams were not included in this study.

The third limitation involves the analysis of the results of the exams. The analysis emphasized the ExamSystem II (ESII) program as used by school division. The ESII program provides basic statistical information for the SC20F exam. The statistical information is compiled by the Assessment and Evaluation Coordinator through data

provided by the science teachers. The statistical information provided by this system is used to analyze the validity and reliability of the SC20F exam and, specifically, to answer the three questions stated above.

Overview of the Study

An assessment of factors related to the fairness and utility of the Grade 10 science standardized exams in the Galileo School Division in Winnipeg, Manitoba, Canada, can be used by school districts to develop policy in assessment and evaluation and valid testing practices. Chapter 1 provides the foundation of the study by declaring the purpose of standardized exams and describing the practices associated with the use of these exams in the Galileo School Division. Standardized exams may be norm-referenced or criterion-referenced with each having a particular purpose and application. Chapter 1 describes the criterion-referenced standardized exams that are used in the Galileo School Division.

Whereas Alfie Kohn (2000) suggests that standardized exams hinder good instruction practices and are inherently biased in their measurements of students' achievement, Richard Phelps (2003) argues that standardized exams provide an invaluable form of accountability and are a positive influence by focusing instruction towards the established curriculum standards. These two opposing viewpoints on standardized exams are representative of two major arguments current in the educational literature. Therefore, chapter 2 is a review of literature associated with standardized testing. Currently, Alfie Kohn, James Popham, E. D. Hirsch, and Richard Phelps are the major protagonists in the debate, and this review examines standardized testing from both a supportive and a contrary perspective. Practically, standardized exams are viewed by the Galileo School Division as a valuable accountability

tool for students, teachers, and administrators that the data from the exam may be used to make informed decisions about instructional practices and divisional policies. But, standardized exams may also create a "narrowing" effect on instruction because teachers may concentrate on the content that will be on the exam rather than teaching towards outcomes more broadly prescribed by the provincial curriculum. Thus, chapter 2 provides an important review of an issue that bears directly on the assessment policy of the Galileo School Division.

Chapter 3 describes the methodology, data, and the analyses of the data. The data consists of the results from the 2006 and 2007 test years, with approximately 1300 hundred students writing the exam. The analyses of this data will determine the reliability and validity of the selected response items. Chapter 3 outlines the goals of the standardized exams as they relate to the research questions. After this, the SC20F exams used in the study are described in detail, as well as the specific procedures used in the invigilation of the exams.

The results of the data analyses are presented in Chapter 4 so that they address the three questions posed in the introduction chapter:

1. Are the SC20F exams in January and June of the same testing year, comparable in purpose and results?
2. Did the change in test development process from the 2006 test year to the 2007 test year affect the quality of the exam and the achievement of students? and
3. Does the difference in the administration and supervision of the exam have an effect on the exam results?

Finally, Chapter 5 concludes the study by discussing the findings in terms of the questions posed in Chapter 1 and the literature reviewed in Chapter 2. Based on the findings and the discussion of these questions, policies for constructing, administering, and correcting the exams will be suggested for the division. Should the practice of standardized exams be continued, should they be continued but with some modifications, or should they be discontinued? Are standardized exams fair or foul in supporting student achievement and measuring student achievement? These are the types of questions that will be discussed in Chapter 5.

CHAPTER 2

REVIEW OF LITERATURE

Standardized testing has long been debated by teachers, students, and parents. The basic questions are: Do standardized tests hinder or enhance the learning of students? Is accountability such an important characteristic in education that it should influence school and divisional policies? Further this, one may be asked, what are the effects on instruction in classrooms that use standardized tests? Supporters of standardized tests note that they "can help teachers deliver effective instruction" (Popham, 2001, p. 15), while those who do not support these exams claim "Teaching to the test narrows the curriculum, forcing teachers and students to concentrate on the memorization of isolated facts, instead of developing fundamental and higher order abilities" (FairTest, 2007).

There are two aspects of standardized tests that must be examined; first, standardization, and second, testing. In this respect, Sireci (2005, p. 113) says: "Standardized simply means that the test content is equivalent across administrations and that the conditions under which the test is administered are the same for all test takers...." Wooldridge (1998) emphasizes the importance of standardized testing beyond the student and classroom level by stating that these tests "allow society to judge the performance of its educational institutions – and to put pressure on them if they seem to be failing in their basic tasks." Moreover, he contends that "For all their imperfections, standardized tests are probably the most powerful instruments of accountability in education". Phelps (2003) identifies two opposing sides in the arguments about standardized testing. On one side, there are the parents, taxpayers, and policy makers who expect an unbiased measurement of the achievement of students, while on the other side, there are teachers and educational

specialists who often claim that a fair assessment of students' achievement can only be conducted by teachers using their own tests. In this chapter, I review the history of testing, and then I examine these two opposing perspectives on the use of standardized tests.

A Brief History of Testing

James Popham (2001, p. 41) notes that on January 22, 1917, during World War I, the German government gave notice that unrestricted submarine warfare would be resumed. After five U.S. vessels were sunk over the next two months, U.S. President, Woodrow Wilson, asked Congress for a declaration of war against Germany and its allies. Immediately, the American government set about mobilizing its military resources, industry, labour, and agriculture, and by the fall of 1917, a U.S. army of over 1,750,000 soldiers had been deployed to France. In deploying these troops, the Army faced a serious dilemma of preparing enough good officers to command this immense fighting force. How were senior officers going to select soldiers so that only those with the "right stuff" were chosen to command the platoons of new recruits?

In 1905, French psychologist Alfred Binet had developed an individual intelligence test. This individual form of testing was not practical for selecting officers from over a million men. Another form of assessment would be needed, and in 1917, a group of psychologists headed by Robert Yerkes from the American Psychological Association developed a test called the "Army Alpha" (Popham, 2001, p. 41). The Alpha contained subtests that assessed recruits' aptitudes in mathematical reasoning, following oral directions, and literacy. The Army Alpha was, in fact, the first group intelligence test developed and used in the United States, and it was composed of multiple choice items. The results for this

test were norm-referenced which enabled senior army officials to differentiate between recruits who were at the 90th percentile from those who were at the 30th percentile in identified aptitudes. Using the rankings of recruits, senior army officers could predict which recruits would be successful officers and which ones would not. The psychologists were very pleased with the technical ease with which the Army Alpha was administered and with the positive results in selecting recruits who would be trained to be officers. In fact, the work on the Army Alpha test helped psychologists develop the variety of standardized tests that exist today.

Types of Tests

The two major forms of tests are criterion-referenced and norm-referenced tests. Both types may be non-standardized or standardized and both may be low-stakes or high-stakes under certain conditions.

Non-Standardized and Standardized Tests

A classroom test created by one teacher for use in his or her course is generally considered to be a non-standardized test because it is constructed, administered, and corrected by the classroom teacher; in other words, the test is unique to that teacher in that particular course at that particular time. Such classroom tests, which are widely used, do not have a standardized format, administration protocol or scoring procedure. Science program reviews in Galileo high schools have revealed that many non-standardized tests are used by science teachers. In these tests, the formats varied considerably, and they included matching items, fill-in-the blank items, selected response items, and free response items. The formats

varied from teacher-to-teacher even though they were teaching the same science courses. Common assessment discussions amongst the teachers who taught the same course, even in the same school, were rare to non-existent. Moreover, the classroom testing conditions and procedures also differed. Seating arrangements, testing instructions, and amount of time given to write the tests varied considerably. In using these tests, teachers would develop scoring rubrics with no standardized protocols. Overall, classroom tests were non-standardized because they differed in format, administration, and scoring practices.

A test is considered standardized if the format, administration, and scoring are consistent regardless of the classroom or school that the students writing the exam attend (Phelps, 2007). The Advanced Placement and International Baccalaureate exams, which are administered world-wide, are examples of standardized tests. Standardization provides a "level playing field" for all students taking the test and avoids undue bias in the exams' format, administration, and scoring.

Criterion-Referenced and Norm-Referenced Tests

Most subject-area tests, such as the science tests used by teachers, are criterion-referenced because they have "items derived from, and weighted in terms of, the objectives of the curriculum, which teachers expect, or should expect, students to achieve" (Cirtwill, Clifton, and Dorsay, 2002, p. 8). The goal of criterion-references tests, of course, is to determine the degree to which the students have mastered the curriculum in a particular subject area. In fact, criterion-referenced tests are most often used to indicate a "mastery" level of understanding based on established standards. In comparison, norm-referenced tests are used to compare how students perform relative to other students, with the average student

being called the "norm" (McMillan, 2007). Obviously, all of the exams analyzed in this study are criterion-referenced tests for which the criteria are specified in the templates for the exams (see Appendix C for an example of a template).

Galileo School Divisions' criterion-referenced exams differ from the classroom tests developed and used by teachers because they are externally developed and graded using specific protocols. In addition, these criterion-referenced tests come with a comparable administrative protocol regardless of the school or teachers of the students. Moreover, the tests are based on standards established from the curriculum outcomes by a group of science teachers, that is, practising grade level and subject matter "experts". In this sense, criterion-referenced standardized tests avoid some of the pit-falls of teacher-created classroom tests. The classroom test may lack proper test construction and permit biased scoring practices. More importantly, classroom tests may not reflect the curriculum carefully enough. One of the responsibilities of the school principal is to ensure proper instruction is taking place but must rely on the expertise of the subject area teacher to fulfill this requirement since the principal is not an expert in all subject areas. In classroom tests, the subject area teacher is considered the expert on the content being assessed with little accountability to students, parents and other citizens who support education by paying taxes. The classroom teacher may choose to provide instruction for only certain topics and criterion-referenced standardized tests would provide a measure of accountability to whether or not the teacher covers all the topics specified in the provincially-mandated curriculum. Specifically, criterion-references standardized tests provide information from a wider perspective than is normal in teacher-created tests, and consequently, they are often superior to teacher-created tests because a number of teachers have worked on them. As such, these tests encourage

teachers to be more sensitive to the requirements of the curriculum and objective in the assessment of the students' mastery of them. Finally, it is often more difficult for students to cheat on a standardized test than on a teacher-created test (Phelps, 2003) because the examination is more formal and often are supervised by a group of teachers/administrators and are seated in a specific seating plan to reduce the opportunities for cheating to occur. The results of criterion based tests are often reported to the students and their parents, but they can be reported at the school-level, the provincial-level, or the national-level. As such, the data may be used to make decisions about individual students, at the school-level, about the effectiveness of instruction and the program at the divisional-level, about the effectiveness of evaluation policy, the provincial-level the effectiveness of the curriculum, or the national-level to assess the level of scientific literacy. As McMillan (2007) says, the test scores should not be the only source of data to make policy decisions, but they should certainly assist in the decisions made regarding instruction and programming, especially at the school level.

In order to ensure that appropriate policies are enacted, McMillan (2007, p. 100) identifies four requirements for interpreting test results:

1. Review the specific standards that were tested, along with sample items, if possible, and know which standards are included in each subscale;
2. Determine the number of items that measure each standard or set of standards;
3. Base your interpretation on how scores are verified by other data or observations; and
4. Be wary of individual item scores, even if aggregated to the entire class, and do not try to analyze each item for each student.

For students, their level of mastery is the percentage correct on the items in the test, and most often the standards for the tests have been determined by groups of teachers teaching the particular subject. In order to have the students do well on the standardized criterion-referenced tests, it is important that the established standards be reasonable for the students and that they are clearly stated. Criterion-referenced tests often contain many items that help to ensure that the test has relatively high reliability and validity coefficients. Many good items, in fact, reduce the scoring bias resulting in a better estimate of each student's level of mastery.

High- and Low-Stakes Tests

According to Popham (2001, p. 33), criterion-referenced tests can be considered high-stakes if two conditions are met: (1) when student performance has significant consequences, and (2) when the scores are indicators of the degree of success of the instructional practices. If students perform poorly on a test, which results in passing or failing a grade, then the evaluation can be considered high-stakes. If the data gathered from students' scores creates a ranking system among schools in a district, which may influence the amount of funding a school receives, then the summative evaluation can also be considered as high-stakes. Low-stakes tests, on the other hand, do not have significant effects on students' final grades; nor do they affect the ranking or funding of schools. Obviously, a quiz given by a teacher would be considered a low-stakes test, but a final exam worth 50% of the students' final grades would be considered a high-stakes test. In his book, "Kill the Messenger", Phelps (2003) expands his discussion on high-stakes tests to include four main benefits, namely, information, motivation, organizational clarity and efficiency,

and goodwill. The benefits of the information category include diagnoses of the mastery level of students, performance accountability of both the school and the school district, and contribute to the selection and recruitment of students' process for colleges and universities. It is arguable that low- and high-stakes tests could include all characteristics, but, the last three, motivation, organizational clarity and efficiency, and goodwill, are intrinsic to high-stakes tests, whereas, low-stakes test likely do not address these issues as strongly, if at all. Motivation can lead to "desirable behaviours, such as students' paying greater attention in their courses and studying more, which, in turn, leads to the accumulation of more knowledge and skill" (Phelps, 2003, p. 226). Motivation not only affects students but can also influence teachers, administrators, school trustees, politicians, and parents. High-stakes tests are, in fact, catalysts for teachers to teach the subject matter well resulting in more efficient instruction and improved teacher-based assessment (Hirsch, 1996). Without such tests, teachers, and administrators are free to teach specific curricula and interpret the curriculum outcomes without concern for any particular standards. Finally, the benefit to high stakes testing is goodwill, contends Phelps (2007). Although difficult to measure, goodwill reflects the public's right to "objective, impartial information about the performance of the public schools' main function – the academic achievement of their children" (Popham, 2003, p. 227). In summary, high stakes tests can have a significant effect on students' final academic standing and promise other benefits for teachers, parents, schools and the public at large.

Advantages and Disadvantages of Standardized Tests

In Fall of 2006, the Galileo School Division published a guide, *A Framework for Thinking About Assessment* (FTA), with the intention to "...support teachers and administrators in implementing the divisional vision for balanced assessment to provide a framework for conversation and thinking about assessment." The guide discusses three forms of assessment; assessment for learning, assessment as learning, and assessment of learning.

Assessment for learning uses strategies that engage students and help them learn more while learning and instruction is going on. The guide includes differentiated instruction to address the area associated with assessment for learning. Referencing the province's education publication, *Success for All Learners* (Manitoba Education, Citizenship and Youth, 1996), FTA suggests differentiated instruction is a strategy for making frequent measurements of understanding – which is a major component of assessment for learning. In the division, FTA relies heavily on Gardner's theory of multiple intelligences as a framework. The FTA describes how a student taking ownership of his learning is a critical component of motivation to learn. The guide refers to the research of Black and William (1998) and concluded "...that successful learning occurs when learners have ownership of their learning..." (Galileo School Division, 2006b, p. 8).

The division's concept of assessment of learning "is about 'looking back' to see how much and how well students have learned at a particular point in time" (Galileo School Division, 2006b, p. 18). The FTA reiterates the different forms that assessment of learning as mentioned in the province's education publications on assessment and communication of learning. These methods include examinations, portfolios, presentations, and performance

assessments. The school division has, in fact, removed standardized exams at grade 3, 7 and 9 in favour of assessments made throughout the year by the teachers, using teacher-created tests such as portfolios, presentations, and exhibitions. The chief superintendent and the school trustees view these forms of assessments as being more authentic than standardized exams. As a result, the science exams may be revised to reflect the division's new assessment policies, which will probably result in a new exam format. Incorporating performance assessments for measuring criterion-referenced curriculum outcomes may be feasible in a single classroom, but may not be readily done at a divisional level due to great difficulty in having all labs standardized in equipment and materials. Phelps (2007) suggests performance based assessments have a greater degree of bias scoring than selected response assessments. In fact, I implemented year-end performance assessments while teaching at a high school in Malawi, Africa, and the amount of equipment required, and preparation time, was enormous. Most science teachers I've spoken to favour performance assessment and use this form of assessment regularly throughout their science courses, but they also recognize the tremendous efforts that would need to be made to insure equality for every student in the standardized performance-based assessment. Nevertheless, the literature on standardized testing has suggested five important questions that should be answered before a school jurisdiction like Galileo School Division modifies or adopts a different assessment and evaluation policy:

1. Do standardized tests support and measure real learning?
2. Do standardized tests narrow the curriculum and promote an antagonistic atmosphere?
3. Is standardized testing a form of reliable assessment?

4. Do stakeholders benefit from standardized tests? and
5. Are standardized tests too costly?

These questions will be discussed from two perspectives, those who opposed standardized testing and those who favour standardized testing.

Do Standardized Tests Support and Measure Real Learning?

Kohn (2000) argues that standardized tests distract from real learning and cannot be used to provide objective assessment. In fact, critics of standardized tests, such as Kohn (2000), often argue that the instructional time consumed in preparing students for the standardized test could be better used by providing more time to their understanding of the course material at a higher level. He argues that teachers are "trapped" in a review process - which typically involves at least a week or more of classes – to help improve their students' test scores by having them practice on old test items. During this time, the students are not exposed to new material or given time to make cross-curricular connections, but are forced to memorize facts, sequential steps, or writing frameworks. Kohn argues that "the test-driven instruction that takes place as a result of accountability-based reform, in U.S. education at least, may reinforce what the worst instructors have been doing" (Kohn, 2000, p. 24). In this respect, the Centre for Research on Evaluation, Standards and Student Testing, states that: "High-stakes testing misdirects instruction even for the basic skills. Under pressure, classroom instruction is increasingly dominated by tasks that resemble tests.... Even in the early grades, students practice finding mistakes [in essays] rather than doing real writing, and they learn to guess by eliminating wrong answers" (Phelps, 2003, p. 43). Kohn notes that the test items on standardized tests examine only what students are able to retain in their

short term memory and are typically knowledge-based questions. In this vein, knowledge level test items fail to examine higher level critical thinking which has the potential of bias in measuring the students' mastery of the subject. If test items attempt to measure higher level thinking, the answer key response typically is an all or nothing response, which is a popular multiple choice response format for standardized tests. Multiple choice items give little credit for students working through the process which is more important than the correct answer. The National Education Association capitulates to Kohn's opinions, "...such tests [multiple choice] do not measure students' ability to analyze, synthesize, draw generalizations, and make applications to new phenomena..." (Hirsch, 1996, p. 179).

Obviously, test scores on free response items do not always indicated the mastery level of the students either. Students scoring well on any test may not have a clear understanding of all the concepts. It is always possible that students may have memorized the steps to arrive at an answer without understand the concept(s) associated with the correct answer. The bias that results from a lack of question objectivity is not restricted to any particular type of test. Often Mathematics, Science, and Social Studies test items tend to be just a regurgitation of isolated facts and English test items examine scripted responses from short creative passages of responses from complete texts. Finally, student behaviour may also have a negative bias on test scores. Kohn (2000) argues that standardized tests create an atmosphere of indifference and test anxiety: "it may be a good proportion of students either couldn't care less about the tests, on the one hand, or care so much that they choke, on the other". Further, William Kilpatrick, founder of Teachers College, states that "the use of such measures [standardized tests] imposes external rewards and punishments for learning rather

than encouraging an inward motivation toward learning for its own sake" (Hirsch, 1996, p. 180).

Kohn appears to support the notion that teachers are the best judges of students when it comes to determining the criteria of the course and the associated nature of instruction and assessment. If teachers were free of the "constraints" placed on them by standardized tests, instructional styles would be diverse and student learning would improve. The core problem with this argument, of course, is that without standards set out by an external agent (e.g. curriculum developers and test constructors), teachers would be free to teach what they wished without fear of accountability. Criterion referenced standardized tests are based on specific learning objectives or essential learnings. As such, the essential learning objectives form a framework from which educators will plan their lessons. Thus, standardized tests do not dictate the nature of instruction, but they require educators to incorporate essential learnings into their lesson plans. Educators would still be able to employ a variety of instruction techniques, such as inquiry learning or design process, in their lessons. Indeed, such techniques are greatly encouraged for science instruction in various states in the United States that also use standardized tests (National Science Teachers Association, 2008). Without standards or essential learnings, teachers would not have a framework to design their lessons and they would be left to develop "standards" which they also deemed important. Typically, these standards would be knowledge based and would not encompass higher levels thinking, such as the application and synthesis of ideas. Essential learnings standards, which form the framework for standardized tests, include both low level understanding and higher level understanding. This way, standardized tests are able to accurately identify the mastery levels of the students by assessing the spectrum of understandings from knowledge

to synthesis to evaluation (Bloom, et al., 1956). Because standardized tests include items from the spectrum of understanding across the various levels, teachers are held accountable for ensuring that their students are also working at the higher levels of understanding. This is not to say there would not be educators who develop similar standards to that of the essential learnings identified by a ministry or division, but we might expect that such cases would be rare.

Educators should be held just as accountable as they hold their own students. If educators are teaching to the test, and the test is based on essential learnings, then they are satisfying the requirements established by curriculum developers. Although educators will teach to the test, and hopefully meet the established standards, it is how they relate the standards to the students which should be the topic of much discussion among researchers and policy makers. Students will experience pressure to do well on standardized tests, but probably no more pressure than what they experience from their teachers' assessments. When students face deadlines, final reports, and numerous tests and quizzes stress is evident. Those students who have a casual approach to standardized tests likely will have the same attitude towards their school work in general. Kohn's argument that standardized tests create an atmosphere of indifference is weak in the sense that the alternative, performance based assessment, does not suggest that this form of assessment is any different from the students' perspectives. Just because the assessment is in the form of a project does not mean that students will demonstrate a higher level of understanding or more enthusiasm. Indeed, performance assessment could be more prejudicial than most critics of multiple-choice exams would like to believe. For example, if the theme of the project, for which certain learning objectives will be measured, involves a topic that the student does not have a

background in or had a bad experience with, that may affect the level of performance of the student. Even if the student has a grasp of the learning objectives, the nature of the theme probably affects his or her ability to demonstrate understanding. For example, in an elementary school, students may be asked to demonstrate their knowledge of life cycles of spiders. Some students may be afraid of spiders, which may affect their ability to convey their level of understanding to their teachers. A high school student is asked to compose an English essay discussing the biased reporting of a particular newspaper article of a sporting event. A student with a strong aptitude for sports may find the sporting theme inviting and be excited to present his thoughts, but a student who is not athletic may not be so "gung ho" composing a paper even though this student may have a strong writing aptitude. In this respect, Hirsch (1996, p. 185) says: "If the topic is especially familiar and congenial to a particular student, the performance is likely to be uncharacteristically good. Conversely, if the topic is unfamiliar, it is likely to be uncharacteristically bad."

Considering that most performance based assessments do not include multiple measurements of the same criteria, a single poor performance often results in a low assessment grade on the essential learnings. Consequently, students have a much better chance of demonstrating their level of understanding on multiple items found in a multiple choice standardized test. The measuring of the performance task brings into question the accuracy of the measurement. A number of studies have shown that the grade received from students' performance-based tests "might well depend more on which year he appeared for the examination, or on which person read his paper, than it would on what he had written" (Hirsh, 1996, p. 184). A fair assessment of a performance-based test can be achieved through multiple assessments of the same performance but only at a considerable cost of time

and money. In this respect, multiple choice standardized tests are often more reliable and efficient measures of real learning than performance-based tests.

Do Standardized Tests Narrow the Curriculum and Promote an Antagonistic Atmosphere?

"Teaching to the test" has become a common complaint of the critics of standardized tests. As one of the most effective critics, Kohn (2000) argues that test items become the content of the curriculum and have a narrowing effect on teaching. First, in some situations, other subject areas may be affected by having their instructional time reduced to allow students to spend more time preparing for tests. In such situations, teachers will not have time to "attend to children's social and moral development – holding class meetings, building a sense of community, allowing time for creative play, developing conflict-resolution skills, and so on – when the only thing that matters is scores on test…" (Kohn, 2000, p. 30). Second, critics argue that teaching to the test may contribute to higher test scores, not because students have a better understanding of the content, but because they have become more familiar with the test. In fact, the critics claim that teaching to the test becomes so important that "what can be measured reliably and validly becomes what is important to know" (Kohn, 2000, p. 35). Third, it is argued that good teachers and administrators are leaving the profession because of the intense pressure associated with standardized testing. As a result, it is becoming increasingly difficult to find qualified people to teach and administer schools. The teachers and administrators remaining become more competitive and defensive, resulting in less and less cooperation and collegial support. Fourth, cheating on high-stakes tests is becoming more frequent. Teachers are succumbing to the pressure from students for higher grades and they are seeking ways, even illegal ones, of increasing

their students' test scores. Finally, educators are beginning to view students with learning difficulties as being a liability because they will lower the average test score of their students, which will affect their reputations, and, in some states in the United States, will affect their salaries.

Proponents of standardized tests, on the other hand, argue that curriculum narrowing is equally prevalent in states and provinces without testing. Teachers can determine the content of the course based on their own personal preference if there are no standardized tests. Phelps (2003, p. 38), for example, argues that "without curriculum accountability, the content of the courses even in the same subject area and grade level can vary widely in content and quality." A student enrolled in a high school physics course, for example, may only be taught mechanics and electricity and may not receive instruction in waves or light because of the physics teacher's personal preference. I have heard evidence of the variability in instruction in the grade eleven physics course in conversations with physics teachers. Occasionally, differences in the curriculum can be observed in the same course taught by different teachers even in the same school. One grade nine science teacher may, for example, focus on only a few topics while another may focus on all the topics in the curriculum. Even if the topics teachers' focus on are similar, the depth of instruction can vary considerably, because of their preferences. The curriculum is not set by the teachers; it is a public document to which the public expects all teachers to show proper defence. In fact, says Phelps, "the public has a legal right to impose curricular order on its schools." "One method by which the public can monitor the implementation of its curricular standards is a standardized testing program" (Phelps, 2003, p. 39). Nevertheless, teaching to the test has been a long-standing criticism of standardized testing.

If teachers are not teaching to the test which is based on curriculum standards, then what are they teaching to? The criterion-referenced standardized tests are based on the essential curriculum standards that all students are expected to achieve. The test does not restrict what instructional practices teachers follow as long as the instruction conveys the curriculum content faithfully. If a teacher chooses to restrict instruction so that it barely covers the content, it is not the measuring tool that is at fault, but the teacher who should be called into questioning. The test does not drive instructional practice; but it is a way to gauge its success. Fortunately, a good educator with solid instructional skills will easily convey standards found on the test without narrowing the curriculum.

Is Standardized Testing a form of Reliable Assessment?

Critics of standardized tests often point out that multiple choice items are to their minds not authentic; that is; they do not measure how well people in the real world understand important things. How often are people required to demonstrate their understanding of a topic by selecting the correct response out of four or five possible choices? Most often, people are asked to demonstrate their understanding by actually performing an activity competently or by engaging in a conversation on a topic. Critics suggest that performance assessments are much more realistic and measure higher-order thinking better than multiple choice test items. In this situation, students do not need to generate responses, but they only need to recognize the best possible response from 4 or 5 alternatives. Moreover, even students who have little knowledge about an issue, may have a good chance of guessing the correct response if they have been properly coached by their teachers. Kohn (2000, p. 11) argues that the multiple choice format does not measure

cognitive process because: "Students are unable to generate a response; all they have to do is recognize one…". To the critics, more authentic performance assessment requires students to perform an activity or create a product (McMillan, 2007, p. 433). Consequently, some teachers assess the performance of their students by using portfolios, laboratory work, and essays, which avoids the problems associated with standardized multiple choice tests.

In response to these critics, Phelps (2003, p. 60) argues that "it is the structure of the question, not the response format that determines the character of the cognitive processing necessary to select a correct answer." Even though the best answer is present in multiple choice items, it may not be easy to determine which response is the best one. So, the real question becomes, can multiple choice items assess both lower and higher levels of understanding? The answer, of course, is "yes". "An ingenious and talented item writer can construct multiple-choice items that require not only the recall of knowledge but also the skills of comprehension, interpretation, application, analysis, or synthesis to arrive at the answer" (Thorndike & Hagen, 1969, p. 103). In fact, Clegg and Cashin (1986, p. 1) say that: "A well designed multiple-choice item can test high levels of student learning, including all six levels of Bloom et al. (1956) taxonomy of cognitive objectives."

Phelps acknowledges that well-designed performance assessment instruments are valid but they often "cost more, take more time to administer, provide less reliable results, take longer to score, and minority students do relatively worse on them" (Phelps, 2003, p. 63). In my position as a divisional science coordinator, I have had many opportunities to discuss performance-based assessment with science teachers in the division, and, in fact, many of them are positive toward this type of assessment, but they agree that the tests are time-consuming to grade. The science teachers also noted that performance assessments

cannot easily be standardized from school to school or even from teacher to teacher in the same school. For example, some schools may not have the proper equipment, and consequently the assessment would, by necessity, be different than in other schools.

One study summarized the authenticity of multiple choice formats by stating: "On the basis of the data examined, we are forced to conclude that constructed response items provide less information in more time at greater cost than do multiple-choice items. This conclusion is surely discouraging to those who feel that constructed response items are more reliable and hence, in some sense, more useful than multiple-choice items. It should be." (Phelps, 2004, p.141)

Do Stakeholders Benefit from Standardized Tests?

The stakeholders associated with standardized testing include teachers, administrators, school trustees, government officials, and of course, students and their parents. Critics of standardized tests argue that publishing school results on standardized tests, even criterion-referenced tests, turn them into the norm-referenced tests. In this respect, Kohn (2000, p. 15) suggests that "norm referenced tests are not about assessing excellence; they are about sorting students (or schools) into winners and losers." He also claims that students experience extreme pressure which causes them to perform poorly.

A report on testing and accountability by the Atlantic Institute for Market Studies notes that several public polls, both nationally and locally, show that the vast majority of parents and citizens favoured "the use of standardized tests for assessing the achievement of students" (Cirtwell, Clifton, & D'Orsay, 2002, p. 10). Students benefit from standardized testing by having clearly defined learning outcomes with instruction focused on helping them

achieve those outcomes. Also, each student will be assessed the same way regardless of the classroom or the school the student attends. It appears that standardized tests are regarded, by parents and taxpayers, as necessary tools for assessing students. World wide, the use of standardized tests has substantially increased over the past two decades (Phelps, 2003, p. 16). The minority of people who are opposed to these tests is, however, very vocal. Local teachers' associations have, for example, come out in opposition to standardized tests. Brian Ardern (2006, p. 2), past-president of the Manitoba's Teachers Society, in an October 2006 article on standardized testing asserted:

> "The results of testing never made any parents more aware of how their kids were doing....[W]e've seen more and more evidence that standardized testing is virtually useless....It has been shown by every available measure that such testing does little or nothing for schools, students or taxpayers....[T]here are flaws in the system, but they won't be fixed by giving more money and attention to standardized testing that does nothing to enhance education or improve the performance of students."

Given the overwhelming support from parents and taxpayers for standardized testing, Phelps (2003, p. 215) asks: Why are teachers' associations often opposed to them? And further, is not the public entitled to know what it expects from public schools? Zwaagstra, Clifton, and Long (2007) suggest that the associations may be afraid that such a testing policy will be used as a form of accountability of teachers and school systems. Furthermore, education policy may be affected by standardized testing programming which may diminish the influence of teachers' union associations. There is little doubt, however, that the vast majority of science teachers in the Galileo School Division favour standardized tests. In fact, many of the teachers use the tests as a framework for their own instruction and assessment.

This practice is even more evident for recently graduated teachers and teachers who are teaching a science for the first time. The science teachers have commented numerous times to me on how the exams help them focus instruction and maintain a timeline to insure that the entire mandated curriculum is taught. Otherwise, many of them would spend more time on their favourite topics and not properly address the entire curriculum. After all, the Province of Manitoba does not have "curriculum police" who insure that teachers deliver the curriculum as it is set out by the Ministry. Consequently, when teachers are provided with a framework in the form of standardized exams, students benefit from the focused instruction. In turn, parents benefit from the knowledge that regardless of the school their child attends, their child will receive instruction at a standard similar to that given other children in other schools in the Division.

Are Standardized Tests too Costly?

Obviously, standardized tests require money for development, implementation, administration, grading, and reporting. Tests are usually constructed by teams of educators and individuals specializing in psychometrics after consulting with appropriate stakeholders, notably trustees, principals, and teachers. The tests are field tested to find errors in the content and the format. During the implementation stage, the tests are copied and packaged, implementation guidelines have to be written and communicated, and the administration of the tests has to be supervised by qualified teachers and administrators. Finally, the tests are scored by qualified teachers. All these stages in the use of standardized tests, of course, cost money which may take resources away from other important educational activities.

Given the size of provincial budgets for education, typically second only to the budget for healthcare, it would be financially irresponsible not to have some sort of accountability system, even at the divisional level, measuring the effectiveness of the education system. Both public and private industry use accountability systems to demonstrate that they are cost effective. Aside from accountability, the cost of using standardized tests far outweighs the operational costs of the system (Phelps, 2003); in fact, the cost of testing students is usually less than one percent of the operational costs. Accepting these costs means that schools are provided with a clear snapshot of how effective they are; they gain reliable and valid information to report to parents and taxpayers; the tests help teachers set their instructional objectives; and students benefit because they become more effective learners. Within the Galileo School Division, the budget set for implementation of standardized examinations, including the writing, reviewing, translation, production, mailing, and marking has been set at $90,000 which is much less than one percent of the school division's budget, about $80,000,000 (Galileo School Division, 2007). In my opinion, this is a very cost-effective tool for measuring the effectiveness of teaching and learning in the schools of the Division.

Summary

A review of the literature on standardized examinations presented in this chapter illustrates that there is a major, continuing and contentious debate about these examinations, with the protagonists disagreeing on a number of points. The fairness of the exams, the narrowing of the curriculum caused by the exams, and whether or not the exams measure real learning are key themes. Nevertheless, most standardized examinations, particularly those

developed and administered by Galileo School Division, are meticulously written, rigorously marked, and fairly reported to students, parents, and citizens, something that cannot be said about classroom tests that have been developed and assessed by individual teachers. Teachers are often extremely busy people with little training in assessment, and consequently, the assessment instruments they develop are not generally very reliable or valid. Moreover, few, if any, teacher-made tests are reviewed by teams of other teachers to eliminate inconsistencies and biases. In this respect, the standardized tests developed and administered have been highly reliable and valid assessment of student's achievement. Importantly, these exams also enable administrators to make informed decisions about educational programming and teaching in individual schools.

CHAPTER 3

METHODOLOGY

The previous chapter examined the contentious arguments for and against standardized tests and assessment practices. The evidence suggests that there is substantial support for standardized tests which reinforces the use of standardized tests as part of its assessment practices in the Galileo School Division. Two types of assessments were identified in the literature, formative and summative assessments. Formative assessment, or "assessment for learning", happens while learning is still underway and includes observations, oral questions, tests constructed by teachers, portfolios, and performances (Davies, 2000). Summative assessment, or "assessment of learning", measures students' achievement after instruction and includes formal tests and examinations. This chapter examines the current practices of summative assessment in the Galileo School Division and focus particularly on the SC20F tests used during the 2006 and 2007 academic years. The chapter focuses on the structure, the implementation, and the grading practices that were used during this time period.

The State of Standardized Tests in Manitoba and in Galileo School Division

In Manitoba, the provincial government mandates that all schools divisions participate in standardized assessments in grades 3, 4, 7, 8, and 12. The Grade 3 assessment focuses on literacy and numeracy; the Grade 4 assessment focuses on French competency of French immersion students; the Grade 7 assessment focuses on mathematics and student engagement; and the Grade 8 assessment focuses on literacy skills. All of these assessments are marked by classroom teachers for their own students using provincially-established

rubrics. The results, of course, are reported to parents and the Department of Education. The two standardized criterion-referenced examinations in Grade 12 are in Mathematics and English, and each has a weight of 30 percent of the students' final grades. Accordingly, these examinations "are intended to provide pertinent information about each student's knowledge and skills in relation to learning outcomes as set out in provincial curriculum documents" (Manitoba Education, Citizenship & Youth, 2007). The exams are marked locally by teams of teachers. In order to have reliable marking, the teachers receive training from the department of education officials on the rubrics and the grading procedures that should be followed.

Galileo School Division, however, is different from other divisions in the province because, in addition to using the provincially-mandated examinations, it also administers its own standardized examinations in Grade 5 and in each grade from 8 to 12, in English, Social Studies, Mathematics, and the Sciences. The examinations range in weight from 20 to 30 per cent of the students' final grades. The examinations are developed by teams of teachers and are marked by classroom teachers who adhere to a scoring rubrics designed by the teams who construct the exams. Test scores are reported to the students and their parents and are reported to the division for use in developing policies; the results, however, are not reported to the government.

Galileo Divisional Standardized Assessments

The Galileo School Division is the only school division in Manitoba to administer standardized examinations, mark them at the divisional level, and publish the results for individual schools. Table 2 describes the grades in which the standardized exams are written

and the subjects that are included. All the exams are created by teams of teachers following guidelines established by the Assessment and Evaluation Coordinator. The first step in the exam development process is identifying the essential learning outcomes of the curriculum in the specific subjects at the specific grade levels, which represent the fundamental concepts and processes that students are expected to know as outlined in the curriculum documents. In my experiences as the science coordinator, the meetings at which teachers are identifying the essential learning outcomes are often quite challenging because teachers may not agree on which outcomes are essential. For example, some teachers may favour a particular topic over others. Nevertheless, the curriculum coordinator has the task of guiding the teachers through the process of identifying the essential learning outcome and seeking consensus among the team members. In some cases, the curriculum coordinator makes the final decisions in order to keep the process moving forward. In these situations, I've accepted that not all teachers will be satisfied, but a decision has to be made, and as the curriculum coordinator, it is my job to do so.

Table 2

Standardized Tests in the Galileo School Division

Grade	Percent of Students' Final Grade	Subjects Examined
5	20 percent	English, Mathematics
8	20 percent	Mathematics, Science
9	25 percent	English, Mathematics
10	25 percent	English, Mathematics, Science
11	25 percent	English, History, Mathematics Biology, Chemistry, Physics
12	30 percent	Biology, Chemistry, Physics

After identifying the specific learning outcomes, the next step is to develop an exam template. Although the exams must consist of both multiple choice items and extended response items, the number of items in each section is at the discretion of the team of teachers constructing the template, see Appendix C for the 2006 SC20F exam template. After the template has been constructed, it is submitted to the Assessment and Evaluation Coordinator for approval, and after it is approved, the curriculum coordinator surveys the teachers who teach the subject for possible items to be included in the exam. Teachers, who are interested in becoming part of the exam development team, inform the curriculum coordinator of their interest, and the coordinator identifies a writer and two exam reviewers. The writer uses the exam template as a guide in developing a draft of the exam, which is then reviewed, using the exam template, sequentially by the two reviewers checking for accuracy, curriculum correspondence, and length.

After completing the task, the first reviewer meets with the curriculum coordinator to discuss possible problems with the exam, and the coordinator takes this reviewer's concerns to the writer. The writer can approve or reject the reviewer's comments, and the curriculum coordinator may override the writer's decisions. In my experiences, the only time that the coordinator overrules a writer is when a question is too difficult. Based on the reviewer's comments and the outcome(s) to be assessed, I will decide to have the item remain or have it rewritten. The review process repeats itself with the second reviewer. After the second review is finished, the curriculum coordinator will review the exam once more. Following this review, the exam is given to the Assessment and Evaluation Coordinator for a final review and for printing. At the time the exam is written by the students, the principal receives enough exams for all the students, which are written at the same time in all schools.

Some schools will administer the exam in a gymnasium, while others will administer the exam in classrooms. In gymnasium settings, a group of teachers supervise the students, while in the classroom setting, only one teacher supervises the students. After the exams are collected, the extended- and selected-response items are corrected by the teachers who taught the course. The exam scores are combined with the teachers' term marks to form the students' final grades. The exam and term marks are reported separately on the students' report cards. The marks and the students' responses for each item on the exam sheets are sent to the Assessment and Evaluation Coordinator. The multiple choice responses are collected from every student who wrote the exam and item analyses are conducted, which are reported to the school principal, the curriculum coordinator, chief superintendent, and the school trustees.

Galileo SC20F Divisional Exam Structure

The Trustees require that the SC20F exam be worth 20 percent of each student's final grade. As noted, the exam is composed of two sections, a free response section and a multiple choice section. The multiple choice section is worth 65 percent of the grade and the free response is worth 35 percent. This study uses the multiple choice section of four SC20F standardized exams which were administered in both January and June of both the 2006 and 2007 academic years.

Table 3 presents information on the four exams that were analyzed in this study. Column two shows that more students wrote the exam in January 2006 and June 2007 resulting from the scheduling at some schools. Column three shows that the exams written in

2006 had 100 items but those written in 2007 had 75 items. The time given to the exam in 2007 however, remained at 2 ½ hours.

Table 3

SC20F Divisional Science Exams for the 2006 and 2007 Testing Years

Exam Date	Student Count	Number of Items	Time Allotment	Selected Response Exam Weighting
January 31 2006	410	100	2.5 hours	65%
June 21 2006	278	100	2.5 hours	65%
January 30 2007	254	75	2.5 hours	65%
June 20 2007	406	75	2.5 hours	65%

The SC20F exam was administered to students in all five high schools in the division, and Table 4 shows the number of students writing an exam arranged by school and testing period. In total, 1348 students wrote the SC20F exam in 2006 and 2007, with about half of the students writing in each year.

Table 4

SC20F 2006 and 2007 Testing Period Data Table

School	School Population	January 2006	June 2006	January 2007	June 2007
School "A"	650-700	99	42	49	72
School "B"	400-450	59	40	45	53
School "C"	850-900	143	76	78	124
School "D"	400-450	21	30	12	54
School "E"	700-750	88	90	70	103
Total		410	278	254	406

The Structure of the SC20F Exam

There are no divisional policies that specify the number of multiple choice items on these exams, but normally there were 100 items. Thus, the SC20F exams in 2006 contained 100 multiple choice items, but the exam in 2007 contained only 75 items. The reduction in the number of items was to address a concern, by the assessment coordinator, that the students who were writing the exam might be becoming fatigued. The multiple choice items were arranged in four sections parallel to the four units of the course: weather, chemistry, physics, and biology. Each multiple choice item contained four possible responses, one of which was better than the other three, and items varied from the knowledge level to the synthesis level throughout the exam. For example, a knowledge level question from the weather unit was: "What is the most abundant gas in the atmosphere? And the possible answers were:

A) hydrogen; B) oxygen; C) nitrogen; D) water vapour." To correctly answer this item, students are required to recall factual information. A higher level thinking item in the physics section was: "As an object begins freefalling, its: A) speed increases. B) acceleration increases. C) A and B. D) none of these." In answering this item, students must recall the concept of gravitational acceleration.

Finally, the multiple choice items had two formats. In the majority of items, a question was asked with the responses in an alternating horizontal format. The following item is from the ecology section of the January 2006 exam illustrates this format:

Which of the following is a density-dependent factor?

A) fire B) earthquake
C) drought D) food supply

The alternative format had an introductory sentence that could be completed correctly by one of the four alternatives written as sentences. The following question, from the weather section of the January 2006 exam, illustrates this format:

During el Nino:

A) the surface temperature of the Pacific Ocean increases, the trade winds travel eastward, and rainfall increase along the coast of Peru.

B) the surface temperature of the Pacific Ocean increases, the trade winds travel westward, and rainfall increase along the coast of Peru.

C) the surface temperature of the Pacific Ocean decreases, the trade winds travel westward, and rainfall increases along the coast of Peru.

D) the surface temperature of the Pacific Ocean increases, the trade winds travel eastward, and rainfall decreases along the coast of Peru.

Galileo School Division Graduate Survey

In October of 2007, the Galileo School Division surveyed 500 graduates of the school division's five high schools. The survey was commissioned to help the Division understand

and meet the needs of the division's students. The purpose of the survey was "to assess recent graduates' perceptions of their high school experience, and to assess to what degree they felt high school had helped prepare them for their present studies or occupation." (Galileo School Division, 2008) The survey was similar to a series of previous graduate surveys, the latest of which was conducted in 2004. The survey of graduates from the Galileo School Division is in keeping with the one of the division's strategic plans; toward an informed community. The survey was conducted by private company hired by the school division through a tender process. The sample consisted of 100 graduates chosen randomly from 2003, 2004, and 2006 and 200 graduates chosen randomly from 2007 for a total sample size of 500 graduates. With 500 – 600 graduates per year on average, the sample size was adequate for statistical purposes. The survey questionnaire was conducted by telephone. Telephone interviews were conducted over the supper hour and call backs were arranged in cases where respondents were not at home for the initial call. Most of the questions in the survey were identical (or almost identical) to those asked in previous years. Therefore, comparisons were made from preceding surveys. The graduates were asked if they benefited from writing divisional exams. The respondents could have replied "strongly agree", "moderately agree", "neither agree or disagree", "moderately disagree", or "strongly disagree". The results of this question in the survey will be discussed in the next chapter.

Galileo SC20F Divisional Exam Administration Procedures

Before the date the science exam was to be written, copies of it were delivered to each school in sealed packages that were not to be opened until the day of the exam. The package contained an affidavit, invigilation instructions, feedback form, bubble sheets for the

multiple choice section, and exam booklets for the free response section. The affidavit was to be signed by the principal of the school acknowledging that the proper procedures were followed in maintaining the security of the exams, and that the pre-established procedures for administering the exam were followed. The exams were scheduled for 2 ½ hours, and the exam supervision was under the control of the school principal resulting in varying types of supervision. Table 5 outlines the various types of supervision used in the five schools.

Table 5

Exam Supervision Formats in the Galileo School Division

High School	Exam Supervision Format
Highschool "B" Highschool "C" Highschool "E'	• Exams are written in school gymnasiums with assigned seating. • Seats are arranged in rows with adjacent rows containing exams from different courses. • Supervision is done by a group of 4-5 teachers including a head examiner who is responsible for taking attendance and announcing instructions. • The groups of teachers are representative of the areas of instruction being tested. • Time reports are given every ½ hour. • Students are allowed to leave after 1 hour of commencement. • Testing group size varies from 75 – 225 students.
Highschool "A"	• Exams are written in classrooms with assigned seating. • Supervision is done by the teacher of the course. • Time reports are given every ½ hour. • Students are allowed to leave after 1 hour. • Testing group size varies from 10 – 30 students.
Highschool "D"	• Exams are written in school cafeteria with assigned seating. • Seats are arranged in table groups with 4-5 students at each table. Each table group has students who are writing two different exams. • Supervision is done by 3-4 teachers. Teachers provide instruction to the students writing exams for their respective courses. • Time reports are given every ½ hour. • Students are allowed to leave after 1 hour. • Testing group size varies from 50 – 125 students.

Following the writing of the exam, the scripts were corrected by using a key, which was developed by the teacher who originally set the exam. The key contained the answers to the multiple choice items and the scoring protocols for the free response items. The students' exams were scored by the teachers who taught the course. The students' bubble sheets for the multiple choice items were marked by the computer at the schools. The totals of the multiple choice and free response were added together to create the final mark for each student which was entered into the teacher's grade book.

All the multiple choice bubble sheets were then forwarded to the Assessment and Evaluation Coordinator to be scored again for the divisional analysis and reporting. The statistical program that did the item analyses computed the following statistics:

1. The difficulty factor for each item;
2. The discrimination index for each item;
3. The Kuder Richardson reliability coefficient for the test; and
4. The mean, median, and standard deviation for the test.

Summary

Galileo School Division is currently the only division in Manitoba to independently use standardized exams as a form of summative assessment of students. The Division uses standardized exams in Grade 5, and in every grade from Grade 8 through Grade 12. The exams assess the core material in the subject of Mathematics, English, Social Studies, and Science. This chapter has outlined the policies and procedures used in assessing the students on the Grade 10 science exams in the Division. Also, it has described the exam development process, which uses subject area teacher teams to construct the exams, which provides the

data from the Grade 10 Science exams administered in January and June of 2006 and 2007 that is used in this study. A sample of the multiple choice items, used in this study, was also presented. In these two years, almost 1400 students wrote the Grade 10 science exams. The results of the analyses of the exams is the subject of the next chapter.

CHAPTER 4

RESULTS

Essentially, this study examined two years of standardized exams in grade 10 science in a Manitoba school division. Surprisingly, this school division is the only one in the Province that uses divisionally-created standardized exams. As such, this study contributes to an understanding of the value of standardized exams and the benefits of having standardized assessment practices. In this respect, Hirsch (1996, p. 214) states: "Without effective monitoring and high incentives, including high-stakes testing programs, no educational system has achieved or could achieve excellence and equity."

A good assessment requires a high degree of validity and reliability. Validity measures the quality it purports to measure and reliability refers to the degree to which test scores are consistent across time, conditions, and test-takers. To insure that a test has a high degree of validity and reliability, the test development process must include:

1. A well defined purpose;
2. Carefully developed test items; and
3. Consistent test administration and scoring procedures

Did the SC20F exams administered in 2006 and 2007 have these characteristics? To answer this question, the data gathered in the study was reviewed and formed the framework for the specific research questions, which help assess the validity and reliability of the SC20F standardized exam:

1. Are the SC20F examinations in January and June, in the same testing year, comparable in purpose and results?

2. Did the change in the test development process from the 2006 test year to the 2007 test year affect the quality of the exam and the achievement of the students? and
3. Does the difference in administration and supervision of the exam have an effect on the exam results?

Are the SC20F Examinations in January and June Comparable in Purpose and Results?

This question is addressed through divisional policy documents and data analysis of exam results. A review of division policies associated with the standardized exams suggest that the purpose or intent of the SC20F exams during the 2006 and 2007 testing periods did not change. The Assessment and Evaluation Coordinator's guidelines associated with the exam did not change nor did divisional policy. That is, the purpose of the standardized exam was to:

1. To provide an additional assessment component, on a wider basis than traditional classroom-based assessments, for individual student evaluation.
2. To provide feedback to staff for reference in the continuous improvement of curricula and instructional practices.
3. To provide information to the general public.
4. To ensure adherence to prescribed curricula.
5. To assist the school division in improving the quality of academic standards.

These guidelines were the same for all SC20F exams written in 2006 and 2007. Thus, in terms of policy, the purpose of the exams was substantially the same for both examinations

periods. Analyzing data from the exams is another method to determine if the exams were comparable in purpose and results.

In the first step of these analyses, the January and June 2006 exams were compared using a reliability coefficient for the two tests. The Kuder-Richardson Formula 20 (KR-20) coefficient measures the reliability of the tests, which is the extent to which they yield the same results on repeated trials (NCS Pearson, 2000). The KR-20 reflects the coherence between the items, and the length of the examination. The coefficients can range from 0 to 1, and high KR-20 coefficients indicate strong relationships between the items in the test. Excellent tests are generally higher than of 0.80 (NCS Pearson, 2000).

Table 6 presents a summary of the KR coefficients for the 2006 and 2007 SC20F exams. The KR-20 coefficient for the January 2006 exam was 0.91 and it was 0.90 for the June 2006 exam. Both exams contained 100 test items. The KR-20 coefficient for the January 2007 SC20F exam was 0.91 and it was 0.90 for the June 2007 SC20F exam, even though only 75 items were used. The Kuder-Richardson 21 (KR-21) coefficient provides a similar indication of the test's reliability, but is slightly lower. The January 2006 exam has a KR-21 coefficient of 0.89 and the June 2006 exam has a KR-21 of 0.87. The January 2007 exam has a KR-21 coefficient of 0.86 and the June 2007 exam has a KR-21 coefficient of 0.84. All of these coefficients suggest that over these two years the tests were highly reliable.

Table 6

KR-20 and KR-21 Coefficients for the 2006 and 2007 SC20F Exams

	January 2006	June 2006	January 2007	June 2007
KR-20	0.91	0.90	0.91	0.90
KR-21	0.89	0.87	0.86	0.84
Number of Test Items	100	100	75	75
Total Number of Respondents	410	278	254	406

Another measure of the similarity of the exams may be indicated by the means and standard deviations of the students who wrote the examination at different times. If these statistics differed substantially, it may indicate that the exams were quite different. The January 2006 exam mean was 63%, with a standard deviation of 14.3, and the June 2006 exam mean was 60%, with a standard deviation of 13.1. The January 2007 exam mean was 63%, with a standard deviation of 14.4, and the June 2007 exam mean was 58%, with a standard deviation of 13.4. The maximum difference between the means was 5%. Consequently, the four exams over these two years seem to be very similar measured in this way.

Two other measures which show the similarity of the exams to be looked at are the pass rates and the distribution of correct response items. Pass rate is defined as scoring 50% or better on an exam. The pass rate for the January 2006 exam was 80.2%, with 410 students writing, and the pass rate for the June 2006 exam was 74.3%, with 278 students writing. The pass rate for the January 2007 exam was 81.0%, with 406 students writing, and the pass rate

for the June 2007 exam was 74.0%, with 254 students writing. The maximum difference between the pass rates within the same testing year was 7%.

Finally, Table 7 shows the distribution of correct answers for the 2006 and 2007 exams. In the two 2006 exams, the correct response was more likely to be "C" and in the two 2007 exams, the correct response was more likely to be 'A" Nevertheless, the similar distribution of correct answers suggests the answer keys for the exams are comparable and minimized any scoring bias.

Table 7

Distribution of Correct Answers for the 2006 and 2007 SC20F Exams

Selected Response Choice	January 2006	June 2006	January 2007	June 2007
A	22%	21%	36%	29%
B	25%	25%	18%	24%
C	31%	32%	25%	24%
D	22%	22%	21%	23%

The degree of difficulty of an item is the proportion of respondents selecting the correct answer. The higher the difficulty coefficient, the greater the proportion of students correctly answered the item. A value of 1.00 indicates that all of the students answered the item correctly, suggesting that, perhaps, the item was too easy. The range of difficulties on good tests depends on what the test administrator wants to measure. If a test is designed to determine if students have mastered a topic, then high difficulty values should be expected, but if the test is to discriminate between students at different levels of proficiency, tests with

difficulty coefficients between 0.3 and 0.7 are most effective (NCS Pearson, 2000). Some authors suggest that the optimal level should be around 0.5 (NCS Pearson, 2000).

Table 8 reports the degree of difficulty for items in the 2006 test year. Not surprisingly, the distributions of the January and June 2006 exams are very similar. In fact, in both tests the great majority of items are in the 0.40 – 0.79 range. In the January 2006 exam, 70 items are in this range and in the June 2006 exam, 62 items are in this range. In the January 2006 test, there were 55 items with a degree of difficulty equal to or greater than 0.60, and in the June 2006 test there were 54 items with a difficulty equal to or greater than 0.60. The average level of difficulty for the items in the January 2006 exam was 0.63 and the average level of difficulty for the items in the June 2006 exam items was 0.60.

Table 8

The Degree of Difficulty for the Item in the January and June 2006 Exams

Degree of Difficulty	January Number of Items	January Percentage of Total Items	June Number of Items	June Percentage of Total Items
0.90 – 1.00	10	10%	6	6%
0.80 – 0.89	9	9%	12	12%
0.70 – 0.79	21	21%	19	19%
0.60 – 0.69	15	15%	17	17%
0.50 – 0.59	22	22%	10	10%
0.40 – 0.49	12	12%	16	16%
0.30 – 0.39	6	6%	13	13%
0.20 – 0.29	3	3%	1	1%
0.00 – 0.19	2	2%	5	5%

Table 9 reports the degree of difficulty for items in the 2007 test year. Not surprisingly, the distributions of the January and June 2007 exams are also very similar. In

fact, in both tests the great majority of items are in the 0.40 – 0.79 range. In the January 2007 exam, 45 items are in this range and in the June 2007 exam, 50 items are in this range. In the January 2007 test, there were 35 items with a degree of difficulty equal to or greater than 0.60, and in the June 2007 test there were 40 items with a difficulty equal to or greater than 0.60. The average level of difficulty for the items in the January 2007 exam was 0.61 and the average level of difficulty for the items in the June 2007 exam items was 0.63.

Table 9

The Degree of Difficulty for the Item in the January and June 2007 Exams

Degree of Difficulty	January Number of Items	January Percentage of Total Items	June Number of Items	June Percentage of Total Items
0.90 – 1.00	4	5%	5	7%
0.80 – 0.89	15	20%	10	13%
0.70 – 0.79	13	17%	17	23%
0.60 – 0.69	3	4%	8	11%
0.50 – 0.59	18	24%	17	23%
0.40 – 0.49	11	15%	8	11%
0.30 – 0.39	6	8%	7	9%
0.20 – 0.29	1	1%	2	3%
0.00 – 0.19	4	5%	1	1%

In 2007, the number of items was reduced from 100 items to 75 items. The possibility of student fatigue in a test of more items was given as the main reason for the reduction in the number of items. If this hypothesis were correct, then we might expect that correct responses for the latter portion of the exam of 100 items may be lower than for those exam items at the beginning of the exam. A level of difficulty analysis of the last 25 items in the 2006 exams and of the last 15 items in the 2007 exams found no such evidence. In both

sets of exams, the level of difficulty distribution of the latter portion of the exams was similar to the level of difficulty distribution for the entire exam. There was no evidence to support the hypothesis that student fatigue influenced exam scores. Table 10 shows a summary of this data.

Table 10

Level of Difficulty for Last 25 Items in 2006 Exam and Last 15 Items in the 2007 Exam.

	Number of Items >0.500	Number of Items Between 0.500-0.059	Number of Items <0.059
January 2006	9/25	7/25	9/25
June 2006	12/25	1/25	12/25
January 2007	4/15	2/15	9/15
June 2007	4/15	1/15	10/15

The SC20F curriculum consists of four topics: ecology, chemistry, physics, and weather dynamics. Table 11 shows the mean scores for the students by topic for each of the exams. The differences across the means are minimal. Although there is almost a significant difference between the means in the chemistry topic for the four exams (~10%), there is little evidence suggesting the exams are not comparable.

Table 11

Means by Topic Area for the SC20F Exams

Exam	Ecology	Chemistry	Physics	Weather Dynamics
January 2006	67.1%	64.2%	62.9%	57.0%
June 2006	64.9%	58.7%	62.4%	53.0%
January 2007	69.8%	55.7%	67.4%	57.0%
June 2007	71.8%	53.3%	65.4%	53.4%

In essence, the first question in this study asked: Is the January and June science exams comparable in purpose and results? The purpose of the exam was to measure the students' mastery level of the grade 10 science content, and the statistical analyses including KR-20 and KR-21 coefficients, divisional exam means, standard deviations, and degree of difficulty measures, were to determine if the exam served this purpose. The statistical analyses showed the results in January and June exams written during both the 2006 and 2007 test periods were very similar. Therefore, it can be confidently concluded that the SC20F exams, from the 2006 and 2007 test periods, were similar in purpose and results. This finding is important given that, for the 2007 testing period, the number of items changed and the exam development process also was changed. Specifically, in 2006, the exam included 100 items and only 75 items in 2007 and a viewing session for interested science teachers was not used in 2007.

Did the Change in Test Development Process from the 2006 Test Year to the 2007 Test Year Affect the Quality of the Exam and the Achievement of the Students?

The development process of the SC20F exams changed from the 2006 year to the 2007 year. In 2006, the test development process involved a viewing session of the draft exam for all interested science teachers, and the majority of teachers participated. The viewing session was held in a divisional meeting centre under the supervision of the divisional Science Coordinator. Copies of the draft exam and the answer key were distributed to the teachers who were permitted to sit at tables and read through the exam items. Teachers were also permitted to discuss the items with each other, but they were not allowed to make notes that they would take out of the room. The draft exams and the answer keys were collected after the viewing session and shredded. Any problems identified by the teachers during the reviewing session were reported to the Science Coordinator for consideration in changing the test. The purpose of the viewing session was twofold: it allowed teachers to check to see if they covered the science curriculum in their teaching, and it helped teachers identify problems with the examination, such as language use and the consonance of the items with the provincial curriculum. This reviewing session process was discontinued in 2007 for apparent security reasons and it has not restarted.

Table 12

Percentage Distribution of Degree of Difficulty for the Items in the 2006 and 2007 Exams

Degree of Difficulty	Percent of Items 2006	Percent of items 2007	Combined 2006/2007 Percentage of items
0.90 – 1.00	8%	6%	7%
0.80 – 0.89	11%	17%	13%
0.70 – 0.79	20%	20%	23%
0.60 – 0.69	16%	7%	11%
0.50 – 0.59	16%	23%	23%
0.40 – 0.49	14%	13%	11%
0.30 – 0.39	10%	9%	9%
0.20 – 0.29	2%	2%	3%
0.00 – 0.19	4%	3%	1%

Table 12 shows a comparison of the degree of difficulty for combined exams in the 2006 and 2007 test years. The 2006 test year had 66% of its test items in the degree of difficulty range between 0.40 – 0.79, and the 2007 test year had 63% of its test items in the same range. Consequently, it seems that the test results are very similar in each year. Given this result and the data provided previously, it can be concluded that the change in the exam development process from the 2006 test year to the 2007 test year did not affect the quality of the exam or the achievement level of the students.

It is also interesting to note that those teachers not involved in the exam development process, either as a setter or reviewer, commonly express concerns regarding the fairness of the exam resulting from the development process. That is, some teachers thought that students in schools which had teachers involved in the exam development process, would score higher than students in schools who did not have teachers involved in the development

process. The data suggests that there is no evidence to support this concern. Each testing period had a setter and two reviewers – each of which was from a different school. As such, there were twelve people who participated in setting and reviewing the four tests. Out of the twelve participants in the exam development process, only five were teachers whose students scored higher than the divisional mean. The greatest difference in the means amongst these five participants was less than 8%. Therefore, there is little evidence suggesting the setter or reviewers involved in the exam development process provided advantages to their students over those teachers who did not participate in the process. The common misconception that the setters and reviewers provide advantages to their students is not restricted to teachers but also includes some administrators. In fact, some administrators have strongly encouraged their teachers to participate in the exam development process to ensure that their students' scores are high. Fortunately, the data does not support these conjectures.

So far, the study questions have focused on data produced at a divisional level, the final question, however, in this study focuses on data produced at the school level and compares the results with respect to how each school invigilated the exams.

Does the Difference Between Schools in Supervision Affect the Results?

As noted, the Assessment and Evaluation Coordinator is responsible for packaging and delivering the exams to the schools. The Assessment and Evaluation Coordinator, in consultation with subject coordinators and school administrators, schedules the date and time for the exam, which universally applies to all students in all the schools. This information is sent to school administrators in the form of a calendar (See Appendix E for 2008 Exam Schedule) that includes the divisional and provincial exam dates and times. The school

administrators are responsible for distributing this information to the teachers and students, and for organizing the invigilation of the examinations. The general guidelines associated with the invigilation include the following (Galileo School Division, 2001):

1. Prior to the scheduled start time of the examination, staff assigned to supervise shall ensure that students are settled and ready to begin.
2. At the scheduled start time, staff assigned to supervise shall distribute the examination papers to students, and instruct them to begin.
3. During the examination, supervising staff members shall restrict their comments to clarification of instructions or procedures only.
4. Under no circumstances shall supervising staff members make any communication – verbal, written, or by gesture or facial expression – which might guide students in determining the answer to a question. This is a matter of common sense and good faith, and supervising teachers must use their good judgment in answering student questions about instructions on the examination.
5. Teachers shall not read the examination to the students. Where, because of a significant handicapping condition of a student who is required to write the examination, it is warranted that the examination should be read to a student orally (or signed, Braille provided, etc.) , this shall be done under the following conditions:
    a. Approval of the Assistant Superintendent, Student Services, is required in advance.
    b. The oral reading of the exam (or signing, Braille, etc.) shall be provided to a single student only, not to a group or class of students.

c. This special arrangement shall be accommodated outside the examination room, so the procedures will not disturb the other students in the class.

6. Should a problem with the examination be discovered by staff during the writing session, this shall be reported immediately AFTER the session, by phone or fax, to the Testing Coordinator. If any adjustment is required, this will be addressed AFTER the writing session, under the direction of the Testing coordinator, through adjustment to the marking procedure.

Nevertheless, the invigilation of the exams varies considerably across the five high schools. Some schools have their students write the exam in a large gymnasium while others have them write the exam in classrooms; some schools have a seating plan, while others have "scramble" seating. As a result, the variability in the administration of the exams poses the third question to be addressed in this thesis: Does the difference in administration and supervision of the exam have an effect on the exam results? Table 13 reports the test means and standard deviations for students in each high school during the 2006 and 2007 testing periods.

Table 13

Mean Scores by High Schools for the 2006 and 2007 Exams

School	January 2006		June 2006		January 2007		June 2007			
	Student Count	$\overline{X}$	Student Count	$\overline{X}$	Student Count	$\overline{X}$	Student Count	$\overline{X}$	Exam $\overline{X}$	S.D.
A	50	58.8	40	59.4	45	58.9	53	61.2	59.6	1.1
B	133	65.8	76	59.0	85	66.9	125	59.9	62.9	4.0
C	88	66.7	91	67.7	70	63.6	106	70.9	67.2	3.0
D	21	47.0	29	51.6	12	54.7	53	52.0	51.3	2.2
E	98	60.9	40	52.9	49	61.1	69	57.3	58.1	3.9

Schools B, C, and E had their students tested in the schools' gymnasium in assigned seating; School A tested the students in the teachers' classrooms in assigned seating; and School D tested the students in the schools' cafeteria at large group tables with random seating. Students were supervised by a small group of teachers in schools A, B, C, and E, and students in school D were supervised by their classroom teacher. The standard deviation from the mean of the schools' exams for all four test periods was very low ranging from 1.1 to 4.0. There is no scoring trend from the January to the June test period in the same test year. That is, students writing in January did not appear to score higher than their counterparts who wrote in June of the same test year. This is also true for the reverse; students in June did not always score higher than those students who wrote in January in the same test year. Overall, it appears that the different invigilation practices did not have an effect on student achievement. Even though the data support this conclusion, the evidence

does not suggest that a standardized invigilation practice should not be used. In fact, the invigilation practice exemplified in schools' B, C, and E should be adopted by the other two schools to avoid any appearance of inappropriate behaviour from the students during an exam writing session.

Galileo School Division Graduate Survey Results

In the fall of 2007, the Galileo School Division commissioned a survey of 500 randomly chosen 2003, 2004, 2006, and 2007 graduates. The respondents were asked if they benefited from the division's standardized exams. In the survey's executive summary, 78.2% of graduates interviewed felt that they benefited from writing examinations (Galileo School Division, 2008). This is a slight increase from 76.2% in 2004 and reflects a gradual increasing trend. Although the respondents were not asked to clarify exactly what the benefits were, the largely positive response from graduates indicates that the standardized exams are viewed as having a positive impact on student learning.

Summary

The study posed three questions to determine the fairness and utility of the grade 10 science exams in Galileo School Division that were administered during a two year testing period in 2006 and 2007. The secondary data analyses showed the four sets of exams administered during this two year period were comparable in purpose and results. Even though the exam development and the invigilation processes varied, the data suggests that this did not significantly affect the performance of the students. Therefore, it appears that the SC20F exams administered during the 2006 and 2007 test years were fair in measuring the

students' mastery regardless of which school they attended. This is very significant because the same conclusion could not be said about classroom-based tests. Teachers are unlikely to use good procedures, such as these, to establish reliability for their own tests, indeed, the reliability and validity achieved in the tests analyzed in this study likely could not have been achieved except by the cooperative, rigorously coordinator effort regulated by a divisional assessment and evaluation policy with clear purposes and protocols.

CHAPTER 5

CONCLUSION

Over the last few years there have been considerable discussions about whether or not standardized exams help or hinder effective teaching and effective learning. The debate has involved many stakeholders, not only teachers and their unions, but also school administrators, school trustees, and academics. These stakeholders have been in a heated controversy about what assessment methods are most effective in improving student learning. This chapter summarizes the findings of the study on the fairness and utility of the Grade 10 science standardized exams in the Galileo School Division, discusses some of the most significant of these findings, and, finally, identifies important implications for educational policy and practice, especially in the Galileo School Division.

Summary

Alfie Kohn (2000) argues that standardized exams hinder good instruction practices and are inherently biased in their assessment of students' achievement while Richard Phelps (2003) argues that standardized exams provide an invaluable form of accountability and are a positive influence by focusing instruction on established curriculum standards. These two opposing viewpoints on standardized exams are representative of one of the most significant of current educational arguments. Upon examination of standardized testing in the Galileo School Division, three questions have been addressed:

1. Are the SC20F exams in January and June of the same testing year, comparable in purpose and results?

2. Did the change in test development process from the 2006 test year to the 2007 test year affect the quality of the exam and the achievement of students?
3. Did the difference in the administration and supervision of the exam have an effect on the exam results?

Currently, E. D. Hirsch, Alfie Kohn, Richard Phelps, and James Popham, are major protagonists in the debate whether to standardize test or not to standardize test. The anti-testing group suggest that standardized tests do not promote or measure real learning but rather encourage rote learning and regurgitation of facts. The proponents also believe that standardized assessment narrows the curriculum and provides few, if any, benefits for students or teachers. Finally, they argue that standardized exams are too costly for school boards which continually face dwindling budgets. On the other hand, the pro-testing community addresses these concerns both logically and empirically and largely dispels them. According to these proponents, standardized testing actually encourages students to learn the mandated curriculum, and as such, help teachers become more effective in their teaching by ensuring that the entire curriculum has been taught. In terms of cost, the Galileo School Division budget for standardized exams is $40,000 per year. In this division, teachers, administrators, students, parents, and trustees consider as a useful and effective accountability instrument that provide data for developing policies about instructional practices.

In Manitoba, the provincial government mandates that all schools divisions participate in standardized assessments in grades 3, 4, 7, 8, and 12. The results are then reported to parents and the Department of Education. These examinations "are intended to provide pertinent information about each student's knowledge and skills in relation to

learning outcomes as set out in provincial curriculum documents" (Manitoba Education, Citizenship and Youth, 2007). The exams are marked locally by teams of teachers, who have received training on the rubrics to be used and the appropriate grading procedures to be followed.

The Galileo School Division is different from other divisions in the province. In addition to using the provincially-mandated examinations, it also administers its own standardized examinations, and these examinations range in weight from 20 to 30 percent of the students' final grades. The examinations are developed by teams of teachers and are marked by classroom teachers who adhere to scoring rubrics designed by the teams who constructed the exams. Test scores are reported to the students and their parents and are reported to the division for use in developing policies. In fact, Galileo School Division is the only division in the province to administer such standardized examinations, mark them at the divisional level, and publish the results for individual schools.

All the exams are created following guidelines established by the Assessment and Evaluation Coordinator. The first step in the exam development process is identifying the essential learning outcomes of the curriculum which represented the fundamental concepts and processes that students are expected to know as outlined in the curriculum documents. After identifying the specific learning outcomes, the next step is to develop an exam template. There are no divisional policies that specify the number of multiple choice items for the divisional standardized exams. For example, in 2007 the science exam contained 75 instead of 100 multiple choice items in an effort to address the possible fatigue of students writing too long an exam. After the template has been constructed, it is submitted to the Assessment and Evaluation Coordinator for approval, and the curriculum coordinator then

assigns teachers to write and review the exam. Following this review, the exam is given back to the Assessment and Evaluation Coordinator for a final review and printing.

The structure of the multiple choice items in the science exams were arranged according to the four topics sections of the course. Each multiple choice item contained four possible responses, one of which was better than the other three, and items varied randomly from the knowledge to the synthesis levels throughout the exam. The multiple choice items had two formats. In the majority of items, a question was asked with the responses in an alternating horizontal format with an introductory sentence that could be completed correctly by one of the four alternatives. The writer used the exam template as a guide in developing a draft of the exam, which was reviewed using the exam template sequentially by the two reviewers checking for accuracy, curriculum correspondence, and length.

Prior to the examination date, the exams were delivered to each school in sealed packages that were not to be opened until the day of the exam. The examination package contained an affidavit, invigilation instructions, feedback form, bubble sheets for the multiple choice section, and exam booklets for the free response section. The affidavit was to be signed by the principal of the school to acknowledge that the proper procedures were followed in maintaining security for the exams. The writing period for the exams was set at 2 ½ hours, and the exam supervision was controlled by the principal in each of the schools. Some principals administered the exam in gymnasia, while others administered the exam in classrooms. In the gymnasia, a group of teachers supervise the students, while in the classrooms the classroom teacher supervises the students.

After the exams were written, the scripts were corrected by using a key that contained the answers to the multiple choice items and the scoring protocols for the free response items.

The students' exams were marked by the teachers who taught the course, and the scores on the multiple choice items and the free response items were added together to create a final exam score for each student.

The exam scores were combined with the teachers' term marks to create the students' final grades. The exam and term marks were reported separately on the students' report cards. The students' multiple choice and free response scores along with the multiple choice bubble sheet responses were then sent to the Assessment and Evaluation Coordinator. All the multiple choice bubble sheets were scored again for divisional analyses and reporting. The statistical program, ExamSystemII, was used to conduct item analyses including computing a difficulty score for each item, a discrimination index for each item, Kuder Richardson reliability coefficients, as well as the mean, median, and standard deviation for all the items in the test. The specific results for each school were reported to the principals and to the divisional administrators. Each school received results for their students and data for all the students in the division, but they did not receive information about other schools. Schools published their exam results in newsletters for parents and on their websites, where anyone could find out the school averages on the exams.

A review of division policies, including the Assessment and Evaluation Coordinator guidelines determined that the purpose of the SC20F exams for the 2006 and 2007 years did not change. The purpose of the standardized exams in the Galileo School Division was:

1. To provide an additional assessment component, on a wider basis than traditional classroom-based assessments, for individual student evaluation.
2. To provide feedback to staff for reference in the continuous improvement of curricula and instructional practices.

3. To provide information to the general public.
4. To ensure adherence to prescribed curricula.
5. To assist the school division in improving the quality of academic standards.

These guidelines were the same for all SC20F exams written in 2006 and 2007. The exams were also compared using elementary statistical analyses, including KR-20 and KR-21 coefficients, exam means, standard deviations, and degree of difficult of the items. The analyses also included the students scoring distribution, distribution of correct answers, student pass rates, and means for the various topics. These analyses showed that the results in all four testing periods were very similar. Therefore, it was confidently concluded that the SC20F exams from the 2006 and 2007 test periods were similar in purpose and results.

In 2006, the test development process included a viewing session of the draft exam for all the science teachers, and the majority of them attended. Copies of the draft exams and the answer keys were distributed to the teachers who were permitted to sit at tables and read the exam items. Teachers were permitted to discuss the items with each other, but they were not allowed to make notes to take out of the room. Any problems identified by the teachers were reported to the Science Coordinator for consideration in revising the test. The viewing session also allowed teachers to determine if they had covered the science curriculum in their own teaching and if the items were congruent with the provincial curriculum. For security reasons, the reviewing process was discontinued in 2007. Nevertheless, the analyses of the multiple choice data showed a similar degree of difficulty for items on the 2006 and the 2007 examinations. It can therefore be concluded that the change in the 2007 exam development process did not affect the quality of the exam or student achievement.

There is an assumption held by both teachers and administrators that teachers not involved in the exam development process, either as writer or reviewer would be at a disadvantage compared to those who were involved. There was also a concern that the students in the schools that had teachers involved in the exam development process would have higher scores than the students in the schools that did not have teachers involved. Fortunately, the results of the analyses did not support this assumption.

The invigilation of the exams varied across the five schools. Some schools had their students write the exam in gymnasia and others had them write in classrooms; some schools had seating plans and others had "scramble" seating. Overall, the results showed that different invigilation practices had no effect on the students' achievement. However, in order to avoid any inappropriate behaviour by students, as observed by some teachers and administrators, supervisory arrangements can affect the students' results, and therefore a standardized invigilation practice should be used in all schools.

Discussion

The purpose of the study was to assess the reliability and validity of the standardized exams in Grade 10 Science courses of the Galileo School Division. Summative assessments, or "assessments of learning", measure how well students have learned the material in the curriculum following a period of instruction and study (Manitoba Education, Citizenship and Youth, 2008). The results of these assessments are used to inform students, parents, teachers, and the public to indicate how well the students are learning the curriculum. The standardized exams that the Division uses must be both reliable and valid in order to be

credible and useful to all those interested in the achievement of the students. In this respect, Manitoba Education, Citizenship, and Youth (2006) states that exams must:

1. Imbed curriculum learning outcomes in the structure and design of evaluation tools;
2. Report student achievement and progress;
3. Inform students and others about progress;
4. Use pre-determined standards or levels of achievement; and
5. Involve parents in discussing learning growth and support; i.e., strengths, areas needing improvements, and goals.

The Galileo School Divisions' standardized exams meet these criteria very well. The divisional exams are criterion-referenced with respect to the provincial curriculum, the exam marks are reported to all stakeholders and made public, and all high schools have active parent councils giving advice on the Division's examination policies.

Obviously, an important characteristic of any assessment instrument is its quality. In this respect, Phelps (2007, p. 89) noted that classroom-based assessments are often hastily constructed by teachers, and rarely, if ever, revised once they have been used. Nevertheless, to insure that a test has a high degree of reliability and validity, the test development process must include:

1. A well defined purpose;
2. Carefully developed test items; and
3. Consistent test administration and scoring procedures.

Kohn (2000) argues that the test items in standardized tests examine only what students are able to retain in their short-term memories and are typically knowledge-based.

As such, knowledge-level items often fail to examine higher levels of critical thinking which create the potential for bias in measuring the students' mastery of subjects. In addition, the behaviour of students during the exam may have a negative effect on their test scores. In this respect, Kohn (2000) argues that standardized tests create an atmosphere of indifference and test anxiety: "it may be a good proportion of students either couldn't care less about the tests, on the one hand, or care so much that they choke, on the other".

Besides being critical of standardized exams, Kohn (2000) supports the notion that teachers are the best judges of students when it comes to determining their mastery of the subjects they teach. If teachers were free of the constraints placed on them by standardized tests, their instruction would be diverse and students would learn considerably more than they do when they are required to write standardized exams. The core problem with this argument is that without standards set out by external agents, teachers would be free to teach what they wished without any accountability. Criterion-referenced standardized tests are based on the specific learning objectives, or essential learning outcomes, of the curriculum. The essential learning provides a framework from which educators plan and teach their lessons. As such, standardized tests do not dictate the nature of instruction, but they require teachers to incorporate the essential objectives into their lessons. Obviously, teachers would still able to use a variety of instruction techniques, such as inquiry learning or design process, but without standardized and examined objectives, they would be left to develop the standards they thought were important. Typically, these standards would be knowledge-based and would not include higher level thinking, such as the application and synthesis of ideas. Essentially, standards provide the framework for standardized tests and include both lower and higher levels of understanding. In this manner, standardized tests are able to

accurately identify the mastery levels of the students by assessing the spectrum of understandings from knowledge to synthesis, and even to evaluation (Bloom et al., 1956).

Not surprisingly, teaching to the test has become a common complaint from the critics of standardized tests. Kohn (2000), once again, argued that tests become the content of the curriculum and have a narrowing effect on the teaching. In his words, teachers will not have time to "attend to children's social and moral development – holding class meetings, building a sense of community, allowing time for creative play, developing conflict-resolution skills, and so on – when the only thing that matters is scores on tests...." (Kohn, 2000, p. 30). The critics claim that teaching to the test becomes so important that "what can be measured reliably and validly becomes what is important to know" (Kohn, 2000, p. 35). In addition, it is argued that good teachers and administrators are leaving the profession because of the intense pressure associated with testing students and reporting the results to parents and citizens. As a result, it is becoming increasingly difficult to find qualified people to teach and administer schools.

Proponents of standardized tests, on the other hand, argued that curriculum narrowing was equally prevalent before testing became a major concern (Phelps, 2003). In fact, Phelps (2003, p. 38) states that "without curriculum accountability, the content of the courses even in the same subject area and grade level can vary widely in content and quality." In fact, in my experience in Manitoba there seems to be considerable variability in the content taught in the grade ten science courses across the school divisions. The curriculum is not set by the teachers; it is a public document to which the public expects all teachers to focus their teaching. As a defender of exams, Phelps states that "the public has a legal right to impose curricular order on its schools." "One method by which the public can monitor the

implementation of its curricular standards is a standardized testing program" (Phelps, 2003, p. 39). Nevertheless, teaching to the test has long been considered a serious criticism of standardized testing. If teachers are not teaching to the test, which is based on curriculum objectives, what are they teaching? The criterion-referenced standardized tests are based on the essential curriculum objectives that all students are expected to know, and the test does not restrict what instructional practices teachers will follow as long as they cover the curriculum content. If teachers choose to restrict their teaching so that they do not cover the content, it is not the measuring instrument, the standardized exam, that is at fault, but the teacher who should be teaching the whole curriculum. The test does not drive instructional practice; it is the instructional practice that drives the test. Fortunately, good teachers with solid instructional skills will easily convey the objectives in the curriculum and potentially found on the test without narrowing the curriculum.

Critics of standardized tests also suggest that multiple choice items are not authentic; that is, they do not reflect how individuals in the real world understand important things. Most often, individuals are asked to demonstrate their understanding by actually performing activities competently, balancing their accounts, driving their cars, or engaging in conversations with other people. The critics argue that these performances should be measured directly because the assessment would obvious be more realistic and it would measure higher-order thinking rather than the lower-order thinking that is typically required by answering multiple choice test items. Kohn (2000, p. 11) argues that the multiple choice format does not measure cognitive process because: "Students are unable to generate a response; all they have to do is recognize one...."

In response to Kohn and other critics, Phelps (2003, p. 60) argues that "it is the structure of the question, not that of the response format, that determines the character of the cognitive processing necessary to reach a correct answer." Even though the best answer is present as multiple choice items, it may not be easy to determine which response is the best one, requiring students to think critically and creatively about the subject. So, the real question becomes: Can multiple choice items assess both lower and higher levels of thinking? Obviously, the answer is "yes." "An ingenious and talented item writer can construct multiple-choice items that require not only the recall of knowledge but also comprehension, interpretation, application, analysis, or synthesis to arrive at the answer" (Thorndike & Hagen, 1969, p. 103). Clegg and Cashin (1986, p. 1) also support this point by saying "A well designed multiple-choice item can test high levels of student learning, including all six levels of Bloom's taxonomy of cognitive objectives."

A number of studies have, in fact, assessed the usefulness of multiple choice tests. One such study stated: "The research evidence...suggests that there is little difference in the knowledge, skills, or abilities measured by multiple-choice and essay (or constructed-response) tests" (Wainer & Thissen, 1993, p. 116) Another study of Advanced Placement tests in seven college subjects concluded that "whatever is being measured by the constructed-response section is measured better by the multiple-choice section" "We have never found any test that is composed of an objectively and subjectively scored section for which this is not true" (Wainer & Thissen, 1993, p. 116).

Also, critics of standardized tests often argue that publishing school results, even for criterion-referenced tests, turn them into the norm-referenced tests. Accordingly, Kohn (2000, p. 15) says that "norm referenced tests are not about assessing excellence; they are

about sorting students (or schools) into winners and losers." He also claims that when students write norm referenced tests, they experience extreme pressure which causes them to perform poorly. Nevertheless, an Atlantic Institute for Market Studies report on testing notes that several public polls show that the vast majority of parents and citizens favour "the use of standardized tests for assessing the achievement of students" (Cirtwell, Clifton, & D'Orsay, 2002, p. 10). Students benefit from standardized tests by having clearly defined learning outcomes with instruction focused on those outcomes. Also, each student will be assessed the same way regardless of the classroom or the school the student attends. Thus, it appears that standardized tests are regarded by parents and taxpayers as necessary tools for assessing students. In fact, around the world the use of standardized tests has substantially increased over the past two decades, with only minority of people being opposed to these tests, but those who are opposed are often very vocal. For example, local teachers' associations have often come out in opposition to standardized tests (Phelps, 2003, p. 16). A former president of the Manitoba Teachers Society has stated (Ardern, 2006, p. 2):

> "The results of testing never made any parents more aware of how their kids were doing....[W]e've seen more and more evidence that standardized testing is virtually useless....It has been shown by every available measure that such testing does little or nothing for schools, students or taxpayers....[T]here are flaws in the system, but they won't be fixed by giving more money and attention to standardized testing that does nothing to enhance education or improve the performance of students."

Given the overwhelming support from parents and taxpayers for standardized testing, Phelps (2003, p. 215) asks two rhetorical questions: "Why are teachers' associations often opposed to them?" "And, is not the public entitled to what it expects from public schools?"

In this respect, Zwaagstra, Clifton, and Long (2007), in another Atlantic Institute for Market Studies report, suggest that teachers' associations may be afraid that such a testing program will create more accountability for teachers and school systems. Moreover, such education policies may diminish the teachers' associations influence. Nevertheless, there is little doubt that the vast majority of science teachers, at least in the Galileo School Division, favour standardized tests. As may be expected, many of these teachers use the tests as a guiding framework for their own instruction and their own assessment, not because they are teaching to the test but because the test represents the objectives for the course. A number of science teachers have commented numerous times to me on how the exams help them focus their teaching and helps them maintain a timeline to insure that the entire mandated curriculum is taught. Consequently, when teachers are provided with a framework in the form of standardized exams, students seem to benefit from the more focused instruction provided by the teacher. In turn, parents benefit from the knowledge that regardless of the school their child attends and the teacher in the classroom, that child will receive instruction that is similar to other children in other classrooms and schools.

Critics also argue that the costs of standardized testing take resources away from other important educational activities. Both public and private organizations, of course, use accountability systems to demonstrate that they can be cost effective. Aside from the importance of accountability in education, the cost of using standardized tests is substantially less than the cost of other educational activities (Phelps, 2003). In fact, the costs of testing students are usually less than one percent of the operational costs of schools. From this expense, schools gain a clearer understanding of how effective they are; they gain reliable

and valid information to report to parents and taxpayers; tests help their teachers set their instructional objectives; and students benefit because they become more effective learners.

It has been shown that the SC20F exams administered during the 2006 and 2007 academic years in the Galileo School Division were comparable in both purpose and results demonstrating an instrument with high reliability. The exam development process ensured a fair exam for students regardless of the classrooms or schools they attended. The development process also ensured that all teachers were aware of the standards and the topics that should have been taught in the course. The test scores revealed that there was no advantage in terms of the students' teacher or school. Administrators in the school division clearly stated that the main purpose of the SC20F exam was to measure the students' mastery of the provincially mandated curriculum. It seems likely that the standardized exams ensure that the entire curriculum was taught and the exam provided a form of accountability to the public. The results were published by each high school and the data was shared with senior administrators and the school trustees to help them make policies. In this way, the results from the tests were used to guide future instructional practices as well as to assist with developing policies and guide school programming. Finally, it was found that the majority of science teachers prefer to have standardized exams to help guide their instruction. The teachers thought that the standardized exams provide useful data to guide their instructional practices, and that the data helped their students focus on the essential outcomes of the curriculum. Consequently, it seems that the divisional standardized exams did not undermine the classroom assessments by teachers; but rather they are required so that students know how well they comprehend the curriculum.

Based upon this research, the practice of standardized exams in the Galileo School Division should be continued to insure the current high standards continue to benefit students, teachers, and parents in the future. Although the standardized exams in the division are very effective, there are a few adjustments that could improve the testing program. These will be discussed in the next section.

Implications

An assessment of factors related to the fairness and utility of the Grade 10 science standardized exams in the Galileo School Division can be used to develop policies to help improve the testing practices. First, the invigilation process should be identical in all of the schools in the division. That is, the environment for which students write an exam in one school should be very similar to the environment in other schools. The most suitable environment to ensure consistency would be to have students write exams in the gymnasia. This would not only ensure a consistent invigilation environment, but it would also discourage cheating that may be easier in the other settings, such as classrooms or cafeteria.

Second, the school division should allow a viewing session for all teachers to read over draft copies of the exams. The viewing session would confirm the teachers' perceptions of the standards to be measured, and it would identify some problems associated with test items. Security concerns could easily be alleviated by imposing strict guidelines for the conduct of the viewing sessions.

Finally, the school division should explore the practice of withholding past exams for student review. The development of standardized exams is very difficult and time consuming, in part, because exam developers must construct different test items for every

examination, all measuring essentially the same learning outcomes. It would be much easier for the creation of tests if the division created a "test bank" of items with thousands of highly reliable and valid items. Exam developers could then use these items in writing specific standardized exams. Periodically, perhaps every three to five years, some test items could be released to teachers for their students to use for testing students in their classrooms. The practice of periodically releasing test items is common in other countries, for example.

In addition, there are some implications of this study for other school jurisdictions and the provincial Department of Education. The study has described in great detail an exam development process that includes how to construct and implement a fair and reasonable standardized examination system. The system developed by the Division was shown to produce reliable exams regardless of the students' classroom, school, or testing year. As such, this system could be used by other schools and divisions in the province. School divisions could, in fact, use these standardized tests as part of their assessment and evaluation practices and the provincial government might find specific ways to assist other schools and school jurisdictions to use them. Such a comprehensive assessment policy, including standardized exams, would ensure more effective teaching and learning in all the provincial schools. Without effective monitoring, good teaching and good educational policies are impossible.

In this respect, the study showed that the Galileo School Division is using a very successful model of developing standardized exams. These exams are, in fact, measuring the student mastery with very high reliability and validity. Regardless of the classroom, school, or testing year, the data showed a comparable mastery level for students. Therefore, all stakeholders, including teachers, administrators, school boards members, parents, and

students can legitimately have confidence that the course is being taught and that the exams are fair and reasonable. What does this mean when it comes to assessments administered by teachers in their own classrooms? A high positive response from recent graduates on the benefit of the divisional standardized exams is a strong indicator of the success of the division's standardized assessment model. The study is not implying, of couse, that all assessments need to be standardized, but that standardized tests have an important place in the assessment of students. Indeed, classroom assessment is important for effective teaching. But assessment practices should reflect a balance between those done by the classroom teacher and those done by the division and the province.

This study also shows that teachers are generally in favour of standardized exams when they participate in the development and the tests are fair and reasonable. It appears that strong teacher involvement in the development process increases their acceptance as being reliable and valid instruments. As such, provincial exams may be similarly appreciated by teachers who become involved in developing provincial examinations. Even though some officials in the teachers' union seem to think that these exams are unreliable and invalid, this opinion is not shared by the science teachers, the parents, or the students, in the Galileo School Division.

REFERENCES

Ardern, B. (2006, October). President's Message. *The Manitoba Teacher*, 85, 2.

Black, P.J., & William, D. (1998). Assessment and Classroom Learning. *Education Assessment: Principles, Policy and Practice, 5*(1), 7-73.

Bishop, J. (1992). Why U.S. Students Need Incentives to Learn. *Educational leadership, 49*(6), 15-18.

Bishop, J. (1994). *The Impact of Curriculum-Based Examinations on Learning in Canadian Secondary Schools.* Ithaca, NY: Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals.* New York, NY: McKay.

Cirtwell, C., Clifton, R., & D'Orsay, J. (2002). *Testing & Accountability: The Keys to Educational Excellence in Atlantic Canada.* Halifax, NS: Atlantic Institute for Market Studies.

Clegg, V.L., & Cashin, W. E. (1986). *Improving Multiple-Choice Tests.* Manhattan, KS: Center for Faculty Evaluation and Development - Kansas State University.

Davies, A. (2000). *Making Classroom Assessment Work.* Courtenay, BC: Connections Publishing.

FairTest (2007, August 20). *How Standardized Testing Damages Education.* Retrieved June 11, 2008, from http://www.fairtest.org/how-standardized-testing-damages education

Galileo School Division. (2000, September 26). *Division Testing Program Policy IL.* Retrieved November 1, 2007, from http://polmanual.sjsd.net/?p=Section%20I%20%20Instruction/

Galileo School Division (2001). *Galileo School Division OP-13005.* Winnipeg, MB: Galileo School Division.

Galileo School Division (2006a). *SC20F January 2006 Examination.* Winnipeg, MB: Galileo School Division.

Galileo School Division (2006b). *A Framework for Thinking About Assessment.* Winnipeg, MB: Galileo School Division.

Galileo School Division (2007). *Galileo School Division Budget 2006-2007.* Winnipeg, MB: Galileo School Division.

Galileo School Division (2008). *Graduate Survey 2007.* Winnipeg, MB: Galileo School Division.

Hirsch, E. D. (1996). *The Schools We Need and Why We Don't Have Them.* New York, NY: Anchor Books.

Kohn, A. (2000). *The Case Against Standardized Testing: Raising the Scores, Ruining the Schools.* Portsmouth, NH: Heinemann Press.

Manitoba Education, Citizenship, and Youth. (2006). *Success for All Learners: A Handbook on Differentiated Instruction.* Winnipeg, MB: Manitoba Education, Citizenship, and Youth.

Manitoba Education, Citizenship, and Youth. (2006). *Rethinking Classroom Assessment with Purpose in Mind.* Winnipeg, MB: Manitoba Education, Citizenship, and Youth.

Manitoba Education, Citizenship, and Youth. (2007a). *Policies and Procedures for Standardized Tests*. Winnipeg, MB: Manitoba Education, Citizenship, and Youth

Manitoba Education, Citizenship, and Youth. (2007b, October 1). *Senior Years English Program*. Retrieved October 12, 2007, from http://www.edu.gov.mb.ca/k12/cur/ela/index.html

Manitoba Education, Citizenship, and Youth. (2008, n.d.). *The Role of Assessment*. Retrieved January 10, 2008, from http://www.edu.gov.mb.ca/k12/assess/role.html

McMillan, J. H. (2007). *Classroom Assessment: Principals and Practice for Effective Standards-Based Instruction*. Boston, MA: Pearson Education, Inc.

National Science Teachers Association (2008, n.d.). *Assessment Practices in Science Classrooms*. Retrieved February 10, 2008, from http://www. nsta.org.

NCS Pearson (2000). *Examsystem II User Guide*. Boston, MA: Pearson Education, Inc.

Phelps, R. (2003). *Kill the Messenger: The War on Standardized Testing*. New Brunswick, NJ: Transaction.

Phelps, R. (2004). *Defending Standardized Tests*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Phelps, R. (2007). *Standardized Testing Primer*. New York, NY: Peter Lang.

Popham, W. J. (2001). *The Truth About Testing: An Educator's Call to Action*. Beauregard St. Alexandria, VA: Association for Supervision and Curriculum Development.

Sireci, S.G. (2005). *The Most Frequently Unasked Questions About Testing*. In R.P. Phelps (Ed.). *Defending standardized testing* (111-122). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Thorndike, R.L, & Hagen, E. (1969). *Measurement and Evaluation in Psychology and Education.* New York, NY: Wiley & Sons.

Traub, R. (1994). *Standardized Testing in Canada; A Survey of Standardized Achievement Testing by Ministries of Education and School Boards*. Toronto, ON: Canadian Education Association.

Wainer, H., & Thissen, D. (1993). Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. *Applied Measurement in Education 6 (2)*: 103-118

Wooldridge, A. (1998, n.d.). *A True Test: In Defense of the SAT.* Retrieved January 15, 2008 from http://www.tnr.com/

Zwaagstra, M., Clifton, R., & Long, J. (2007). *Getting the Fox Out of the Schoolhouse: How the Public Can Take Back Public Education.* Halifax, NS: Atlantic Institute for Market Studies.

Appendix A

GALILEO SCHOOL DIVISION 2007-2008 Exam Guidelines

GUIDELINES FOR EXAM COMMITTEE MEMBERS 2007-2008

Coordinators, setters, reviewers, and translators, thank you for taking the responsibility to participate in the divisional assessment development for the 2007-2008 school year. Please read all guidelines to ensure that processes are followed. That is, if you are reviewer, please read the setter guidelines as well, etc. to develop a broad understanding of the exam development process.

Thank you.

EXAM SETTERS

The following guidelines are to be considered when preparing divisional exams:

1. Security of the examination, and any parts thereof, must be strictly maintained. Any exam hard copy, CD, or disk, or any other format, must be maintained in a locked storage area at all times. The intent of this guideline must be maintained at all times. The exam, in part or in whole, draft or completed, may not be shared with anyone, in any form.

2. The Division core curriculum or MB. Education guidelines are to be adhered to.

3. All levels of thinking skills are to be considered.

4. Exams should include 10% of their value based on higher level thinking questions. This does **not** mean questions from the next grade's curriculum, but rather refers to questions based on the current grade level curriculum outcomes.

5. The target length should be an exam for which students will require about 15 or 20 minutes **less** than the prescribed time for the exam for completion.

6. Exam Committees should consult previous exam item analyses to guide construction of items which meet the targets set out in items 3 and 4. Make the changes required to ensure that #3 and 4 and adhered to.

7. Exam Committees should ensure that any multiple choice items have a corresponding list of curriculum outcomes, which correlate to each of the test items.

8. Most multiple choice items should be written in the positive, rather than the negative i.e., which is the **best**..... which is an example of..... , which of the following is **true** (rather than "which is **not** an example of".... or which is false).

9. Correct answers on multiple-choice questions should not follow a pattern.

10. A subtest should contain no fewer than 6 questions to be statistically valid. There can be no more than 8 subtests in an exam excluding the extended, process or open response section of the exam.

11. For High School exams, which are being written in both January and June, the basic format and weighting of sections should remain the same from one semester to the other within the same school year.

12. Teacher comments from the previous exam are to be considered. The coordinator will have a copy of teacher comments regarding the pertinent exam.

13. Exams submitted to the Assessment and Evaluation Coordinator should conform to the general appearance of all Division exams. Please submit a hard copy **and** CD.

14. Diagrams, graphs, pictures, maps, etc. must be submitted in a manner that allows them to be scanned into the exam without revision.

15. Multiple choice keys, extended answer marking guides, process/open-ended rubrics, examination Information Bulletin, and back-up questions must be submitted with the exam (see #15).

16. **When preparing multiple choice questions, additional questions must be prepared** to form the basis of a back-up exam which will be used in case of a breach of security, a student writing at a time outside the scheduled exam period, etc.. The additional/back-up questions (for the multiple choice section of the exam only) should amount to **20% of the multiple choice** questions on the exam. These questions should be submitted with the actual exam and should be given the number of the exam question which each would replace if it was used. The previous June exam may be used as the source for the 20% extra multiple choice questions required for the back-up exam. Just circle questions to be used as back-up questions and indicate which question each would replace on the current exam.

17. Information Bulletin: Setters must submit a reasonably detailed outline of exam content (examples of questions if necessary) and exam weights (the distribution of questions on the exam based on the various units) must be submitted with the exam to the Assessment and Evaluation Coordinator for distribution to all

teachers whose classes will be writing that exam. People need a reasonable amount of information as to what to expect. The content of the exam should come as a surprise to no one. Upon request, a copy of the previous exam and Information Bulletin will be made available to the Exam Committee.

18. All Division exams will contain an extended answer, process or unstructured response segment. **The proportion of the exam represented by this segment must not change from the previous year's exam without consultation with the Subject Area and Assessment and Evaluation Coordinators**.

19. Division exams will count for 20% at Gr. 5, 6, 7 and 8; 25% at Gr.9, 10 and 11, and 30% at Gr.12.

20. DEADLINES:
    - 1st Sem. exams for High School are to be submitted to the Assessment and Evaluation Coordinator **no later than October 31**; earlier is preferred.
    - Early Years and Middle School exams are to be submitted to the Assessment and Evaluation Coordinator **before the Christmas break**.
    - 2nd Semester exams for High School are to be submitted **no later than January 31**.

21. Exam Delivery or Pick-up: Please do not use the Divisional courier for exams. Once your exam is ready, a) give ASSESSMENT AND EVALUATION Coordinator a call at 999-9999 to pick it up or 2) have the Subject Area Coordinator deliver it to the ASSESSMENT AND EVALUATION Coordinator.

22. **EXAM REVIEWERS**: Security of the examination, and any parts thereof, must be strictly maintained. Any exam hard copy, CD, or disk, or any other format, must be maintained in a locked storage area at all times. The intent of this guideline must be maintained at all times. No part of the exam may be shared with anyone, in any form, including email. The exam, in part or in whole, draft or completed, may not be shared with anyone, in any form.

    When the exam is ready, it will be delivered to you for review. Please complete the review **within a one week timeline**. Please do **not** use the Divisional courier for exams. Call or email the ASSESSMENT AND EVALUATION Coordinator or the Subject Area Coordinator when ready for discussion of the proposed changes and exam pick-up. The exam will be returned to the setter for consideration of the reviewer's recommendations. Once changes have been made, the exam will be delivered to a second reviewer for final review and proposed changes.

Both reviewers are asked to check exams by actually "writing" the exam as a student would and checking for:

A) accuracy of content
B) accuracy, thoroughness and appropriateness of keys/marking guides/rubrics
C) questions on material that may not be on the curriculum. PLEASE NOTE: **For any unsuitable question, include a new question for consideration.**
D) clarity of wording
E) formatting problems (are things labelled correctly and clearly, has sufficient space been left for students to do the work, etc.)
F) typos (spelling, omissions, some multiple choice items that say A.B.B.D. instead of A.B.C.D., etc.)
G) Multiple Choice question stems in the positive versus negative (see #15 above)
H) Length of exam - a recommended procedure is that the reviewer write the exam as if he/she were a student and then multiply his/her time by about 2.5 to get an estimate of the length of time an average student is likely to require. Remember that the goal is to have an exam which will take the average student 15 or 20 minutes less than the time allotted for the exam.
I) accuracy of the answer key (bubble sheet and extended/long answers) and/or marking guide

Any concerns arising from the above directions should be discussed with the Assessment and Evaluation Coordinator.

Appendix B

PART B: MULTIPLE CHOICE – 100 questions (65% of exam mark)

I: Dynamics of Ecosystems

1. The ultimate source of energy in MOST ecosystems is the

   A) plants.
   B) sun.
   C) wind.
   D) cosmic forces.

2. In the water cycle, MOST water is lost from oceans through

   A) respiration.
   B) transpiration.
   C) evaporation.
   D) photosynthesis.

3. Part of every ecosystem includes the living or _____factors.

   A) biotic
   B) abiotic
   C) detritus
   D) transgenic

4. As the population reaches the carrying capacity, which of the following is predicted by the logistic growth curve?

   A) Population density will increase exponentially.
   B) Population density will decrease exponentially.
   C) Population growth rate will increase.
   D) Population growth rate will level off.

5. Which of the following is a density-dependent factor?

   A) fire
   B) earthquake
   C) drought
   D) food supply

6. All individuals of the same species within a given area comprise a

   A) population.
   B) abiotic factor.
   C) community.
   D) ecosystem.

7. The role an organism plays within its community is its

   A) habitat.
   B) niche.
   C) symbiosis.
   D) carrying capacity.

8. Another term for a producer is a/an

A) heterotroph.
B) autotroph.
C) zygomorph.
D) consumer.

9. Interspecific competition occurs between

A) members of the same species.
B) members of different species.
C) different communities.
D) tribes.

10. What would have a positive effect (i.e., increase) on a population?

A) increased natality
B) increased mortality
C) decreased immigration
D) increased emigration

11. Which of the following is considered a density-independent factor for population growth?

A) waste accumulation
B) food supply
C) flood
D) mating opportunity

12. Populations tend to increase in numbers until their environment can no longer support their demands. The maximum number of individuals that can be supported in an environment is known as the

A) logistic growth.
B) carrying capacity.
C) sustainable development.
D) biodiversity index.

13. In the water cycle, plants return water to the atmosphere by

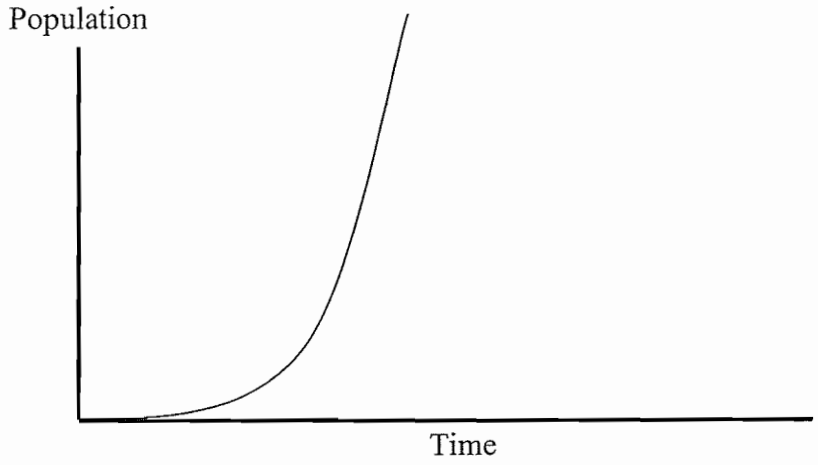
A) transpiration.
B) evaporation.
C) respiration.
D) condensation.

14. Since the beginning of the Industrial Revolution humans have used carbon based sources for combustion to drive machinery. What consequences might this have had on the carbon cycle?

A) The amount of CO_2 released into the atmosphere has remained the same.
B) The amount of CO_2 released into the atmosphere has increased.
C) Oceans are able to dissolve more CO_2.
D) The amount of CO_2 released into the atmosphere has decreased.

15. The following growth curve is also called a/n _____ curve.

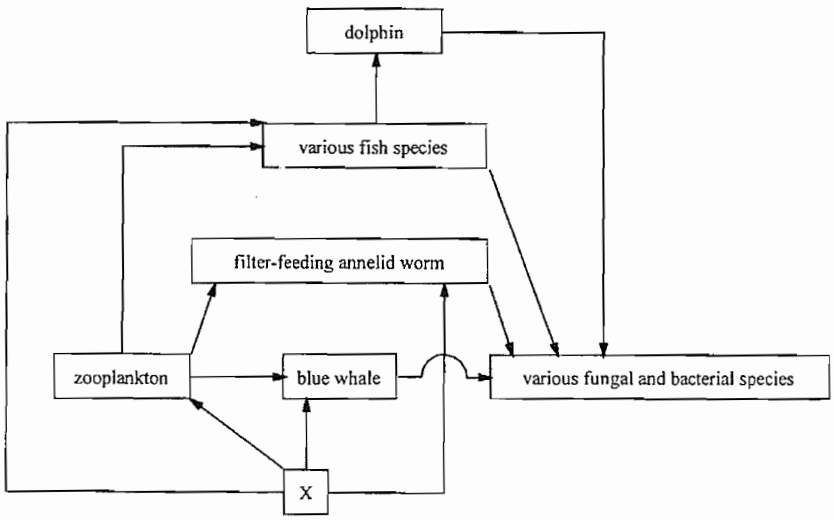
A) logistic
B) exponential
C) "S" –shaped curve
D) sigmoidal



16. What describes the movement of energy and nutrients in a typical ecosystem?

A) Both energy and nutrients are recycled.
B) Energy can be recycled but nutrients cannot.
C) Energy cannot be recycled but nutrients can.
D) Neither energy nor nutrients are recycled.

The diagram below shows a food web. It refers to question #17.



17. What is species X's role in the ecosystem?

A) a decomposer
B) a producer
C) a primary consumer
D) a secondary consumer

18. Food webs are more realistic models than food chains because food webs describe

A) an ecosystem.
B) abiotic factors.
C) more interactions.
D) none of these.

19. Which chemical cycle includes lightning and bacteria?

A) carbon
B) water
C) nitrogen
D) phosphate

20. A herbivore would occupy which trophic level?

A) first trophic level
B) second trophic level
C) third trophic level
D) more information needed

21. An example of an animal in the third trophic level would be

A) a herbivore.
B) a carnivore.
C) a cow.
D) a chicken.

22. Ecology is the study of

A) the environment.
B) pollution.
C) interspecies relationships.
D) relationships between organisms and their environment.

23. If a community consisting of a complex food web were reduced to a single food chain, the result would be

A) more stability.
B) less stability.
C) no effect on stability.
D) increased biodiversity.

24. In a pyramid of biomass for a terrestrial biome, which level would represent the greatest overall biomass?

A) producers
B) primary consumers
C) top carnivores
D) decomposers

25. In the carbon cycle, animals and plants return carbon to the atmosphere by

A) fermentation.
B) photosynthesis.
C) respiration.
D) combustion.

II: Chemistry in Action

26. Which of the following families of elements is considered the MOST reactive?

A) alkali metals
B) alkaline earth metals
C) chalcogen
D) noble gases

27. Which of the following elements is a halogen?

A) hydrogen
B) lithium
C) oxygen
D) fluorine

28. Alkaline Earth metals have

A) 1 valence electron.
B) 2 valence electrons.
C) 3 valence electrons.
D) 7 valence electrons.

29. Which statement about the Law of Conservation of Mass is **TRUE**?

A) During a chemical reaction, such as the burning of wood, the products of the reaction can have a mass less than the reactants in a closed system.
B) The reactants in a chemical reaction have a different mass than their products.
C) The mass of the reactants and the mass of the products in a chemical reaction will always be the same in a closed system.
D) None of these statements is true.

30. Which of the following is NOT a balanced equation?

A) $2HCl + 2Na \rightarrow H_2 + 2NaCl$
B) $MgF_2 + NaCl \rightarrow 2NaF + MgCl_2$
C) $Mg(OH)_2 + 2HCl \rightarrow MgCl_2 + 2H_2O$
D) $2C_2H_2 + 5O_2 \rightarrow 2H_2O + 4CO_2$

31. In general, when substances undergo a chemical change, they

A) retain all their original properties.
B) produce a mixture.
C) produce new substances with new properties.
D) remain the same substances, but with new properties.

32. Elements with 5,6, or 7 electrons in their outer shell tend to be

A) metals.
B) inert(noble) gases.
C) metalloids.
D) non-metals.

33. Predict the results of adding carbon to oxygen.

A) An ionic bond would form.
B) A covalent bond would form.
C) The product would weigh less than the reactants.
D) The product would weigh more than the reactants.

34. An Iron III ion has a combining capacity of

A) 3+
B) 3–
C) 3+ or 3–
D) 5–

35. Of the following pairs of elements, _____will form covalent bonds (instead of ionic bonds).

A) potassium with aluminium
B) carbon with oxygen
C) sodium with oxygen
D) calcium with iodine

36. Negative ions are formed from neutral atoms by a

A) gain in neutrons.
B) gain of electrons.
C) loss of protons.
D) loss of neutrons.

39. Reactions in which two elements or compounds unite to form one new product are called

A) single displacement reactions.
B) double displacement reactions.
C) decomposition reactions.
D) synthesis reactions.

40. Dicarbon hexahydride would have a molecular formula:

A) C_2H_8
B) C_6H_2
C) C_2H_6
D) CH_6

41. Which of these equations is correctly balanced?

A) $3Fe + 2CuCl_2 \longrightarrow 3FeCl_3 + 2Cu$
B) $3Fe + 2CuCl_2 \longrightarrow 2FeCl_3 + 2Cu$
C) $3Fe + CuCl_2 \longrightarrow 3FeCl_3 + Cu$
D) $2Fe + 3CuCl_2 \longrightarrow 2FeCl_3 + 3Cu$

42. Which of these is a single displacement reaction?

A) $Fe + CuCl_2 \longrightarrow FeCl_3 + Cu$
B) $NaOH + HCl \longrightarrow NaCl + H_2O$
C) $H_2O_2 \longrightarrow H_2O + O_2$
D) $C_6H_{12}O_6 + O_2 \longrightarrow CO_2 + H_2O$

43. Acids are substances which form_____ when placed in water.

A) hydrogen (H^+) ions
B) hydroxyl (OH^-)
C) salt ions
D) chloride ions

44. If a substance has a pH = 14, it would be a

A) strong acid.
B) weak acid.
C) strong base.
D) weak base.

45. Sharing of electrons occurs in

A) the formation of ions.
B) covalent (molecular) bonding.
C) ionic bonding.
D) a salt (NaCl) molecule.

46. Which of the following BEST represents a decomposition reaction?

A) $AB + CD \longrightarrow AC + BD$
B) $A + B \longrightarrow AB$
C) $AB \longrightarrow A + B$
D) $A + BC \longrightarrow AB + C$

47. A chemical reaction in which elements exchange "partners" can be considered a

A) synthesis reaction.
B) single-replacement reaction.
C) decomposition reaction.
D) double-replacement reaction.

48. All acids contain the element

A) oxygen.
B) hydrogen.
C) nitrogen.
D) chlorine.

49. What is the TRUE effect of a base on an indicator?

A) Blue litmus paper turns red.
B) Phenolphalein stays colourless.
C) Blue litmus paper turns yellow.
D) Phenolphalein turns pink.

50. A neutralization reaction is one that

A) can create a new acid.
B) can be identified by its sour taste.
C) turns red litmus paper blue.
D) produces water and a salt.

III: In Motion

51. A scalar quantity has all the following except

A) size.
B) direction.
C) unit.
D) numerical value.

52. Which of the following is an example of displacement?

A) 40 km
B) 20 km/h[E]
C) 1.5 m [right]
D) 15 km/h

53. Isaac runs 5 km [W], 10 km [E] and then 15 km [W]. If east is positive, what is resultant displacement?

A) +10 km
B) 30 km
C) –10 km
D) +15 km

54. Annie walks 2.5 km [E], 3.5 km [S], 3.5 km [W], 1.0 km [E] and then 3.5 km [N] in 20 hours. What is Annie's average velocity?

A) zero
B) 14.0 km/h [N]
C) 0.7 km/h
D) 2.0 km/h [S]

55. An object moves at a constant rate from 0 m to 20 m in a straight line. If it took 4 seconds, calculate the object's speed.

A) 80 m/s
B) 16 m/s
C) 5 m/s
D) 0.2 m/s

56. Which statement <u>BEST</u> describes the relationship between distance travelled (of an unrestrained passenger in a collision) and velocity?

A) Distance is proportional to velocity.
B) Distance is proportional to velocity squared.
C) Distance is not proportional to velocity.
D) Distance is inversely proportional to velocity.

57. The values needed to calculate the velocity of an object are

A) speed and time.
B) distance and time.
C) distance, direction, and time.
D) acceleration and speed.

58. Acceleration is defined as

A) change in position over a period of time.
B) change in velocity over a period of time.
C) the same position over a period of time.
D) the time it takes for an object to go from position 1 to position 2.

59. Aristotle proposed that

A) force is proportional to acceleration.
B) force is proportional to mass.
C) force is a natural motion towards the centre of the universe (Earth).
D) forces do not always occur in pairs.

60. Newton's First Law states that

A) an object will remain at rest or in motion until a force acts on it.
B) action and reaction are equal and opposite.
C) an object's force is the product of the mass of the object and the acceleration of its motion.
D) the rate of change of momentum is proportional to the imposed force, and goes in the direction of the force.

61. Which of the following will affect the braking distance of a vehicle?

A) condition of the driver
B) condition of the road
C) speed of the vehicle
D) All of these choices are possible.

62. What is 35.0 km/h equal to in m/s?

A) 0.035 m/s
B) 126 m/s
C) 583 m/s
D) 9.72 m/s

63. The slope of a velocity-time graph will determine the

A) speed of the object.
B) distance of the object.
C) acceleration of the object.
D) displacement of the object.

64. A vehicle has to stop quickly. If it travels 20 m and the speed is 2 m/s, what is the coefficient of friction (k) on this surface?

A) 10
B) 4
C) 5
D) 40

65. A skateboarder travels 30 meters in 10 seconds at a constant rate. If k = 2, how much distance does the skateboarder need to stop?

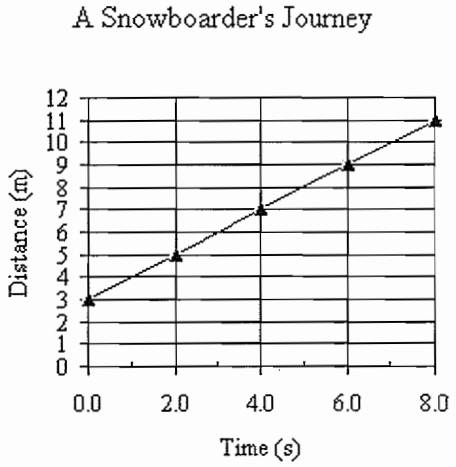
A) 18 m
B) 10 m
C) 15 m
D) 150 m

66. A zero slope on a distance-time graph indicates

A) the object is not moving.
B) the object's speed is increasing.
C) the object has a low speed.
D) the object has a high speed.

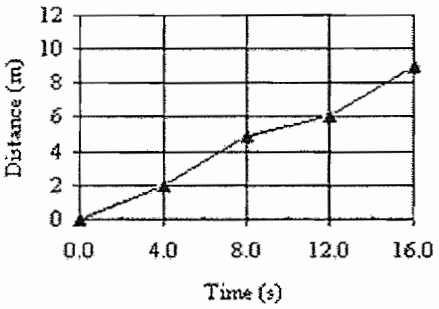
67. Determine the slope of the following graph.

A) 2.5 m/s
B) 3.0 m/s
C) 1.0 m/s
D) 1.4 m/s

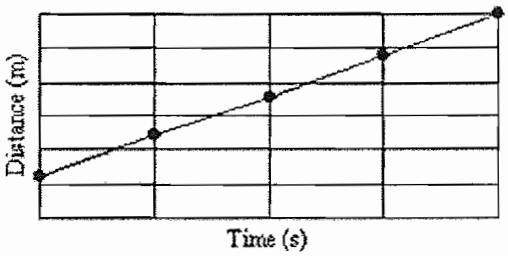


68. Which of the following graphs does NOT represent an object travelling at a uniform speed?

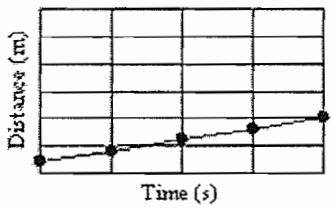
A)



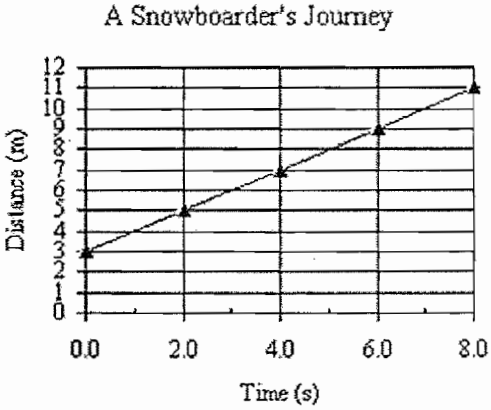
B)



C)

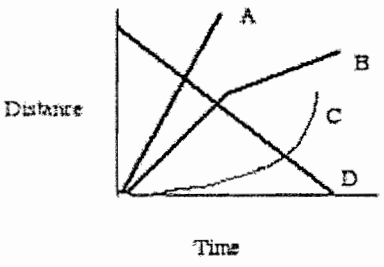


D)



69. Which of the following graphs illustrates a constant negative velocity for the whole trip?

A) A
B) B
C) C
D) D



70. A person kicks a soccer ball from rest to a velocity of 20 m/s [toward the goal] in 0.5 s. The acceleration of the ball is

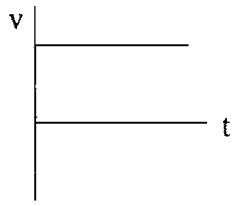
A) 20 m/s^2 [away from the goal].
B) 0.025 m/s^2 [toward the goal].
C) 40 m/s^2 [toward the goal].
D) 10 m/s^2 [away from the goal].

71. What amount of time is required for an object to cover a distance of 21.98 m if it is travelling at 1.27 m/s?

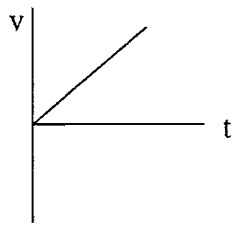
A) 380.6 s
B) 17.3 s
C) 27.9 s
D) 0.06 s

72. Which of the following velocity graphs represents an object that is slowing down?

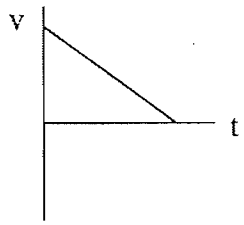
A)



B)



C)



D) None of these.

73. Considering both reaction time and braking time, which of the following would NOT be a factor in the stopping distance of a moving car?

A) tire traction of the car
B) mass of the car
C) speed of the car
D) make and model of the car

74. Which of the following is NOT a condition of acceleration for a car that is accelerating?

A) moving with an increasing speed
B) moving with a decreasing speed
C) moving with a high speed
D) changing direction

75. As an object begins freefalling, its

A) speed increases.
B) acceleration increases.
C) A and B
D) none of these

IV: Weather Dynamics

76. The MOST abundant gas in the atmosphere is

A) hydrogen.
B) oxygen.
C) nitrogen.
D) water vapour.

77. Which is the CORRECT order of the atmospheric layers?

A) mesosphere, stratosphere, exosphere, tropopause
B) salt water, fresh water, troposphere, stratosphere
C) fresh water, troposphere, stratosphere, mesosphere
D) troposphere, stratosphere, mesosphere, thermosphere

78. Which part of the atmosphere is responsible for <u>most</u> weather systems?

A) troposphere
B) stratosphere
C) mesophere
D) hydrosphere

79. On the Earth, 50% of the solar radiation is absorbed by

A) the air.
B) clouds.
C) Earth's surface.
D) the atmosphere.

80. Atmospheric pressure is

A) the pressure of atmospheric gases due to gravity.
B) the pressure of water and atmospheric gases.
C) the surface to space pressure of gases.
D) a value that can be measured with a thermometer.

81. As air pressure increases, air density

A) decreases.
B) increases.
C) stays the same.
D) not enough information

82. The method by which energy is transferred from the Sun to Earth through outer space is

A) conduction.
B) convection.
C) radiation.
D) all of the above.

83. Which of the following is listed in order from the highest to the lowest albedo?

A) coal; a glacier; an ocean
B) an ocean; coal; a glacier
C) a glacier; coal; an ocean
D) a glacier; an ocean; coal

84. Which of the following statements is TRUE about the hydrosphere?

A) It is made up of fresh water, salt water, and glaciers.
B) It describes 30% of Earth's water is fresh water.
C) No new water is produced anywhere on Earth.
D) 99% of Earth's water is salt water.

85. The study of the Earth's atmosphere and weather forecasting is called

A) anemology.
B) astronomy.
C) meteorology.
D) morseology.

86. When a cold dry current, around 30° latitude, reaches the west coast of a continent, it tends to produce

A) a temperate rainforest.
B) a tropical rainforest.
C) a desert.
D) grassland.

87. Frontal clouds/weather fronts are formed when

A) the leading edge of a large moving mass of air meets another mass at a different temperature.
B) air near the ground absorbs energy from a heated surface such as a lake, asphalt, or dirt.
C) aluminium sulphate is artificially introduced into the atmosphere.
D) air moves up a mountainside and expands and cools because of lower pressure.

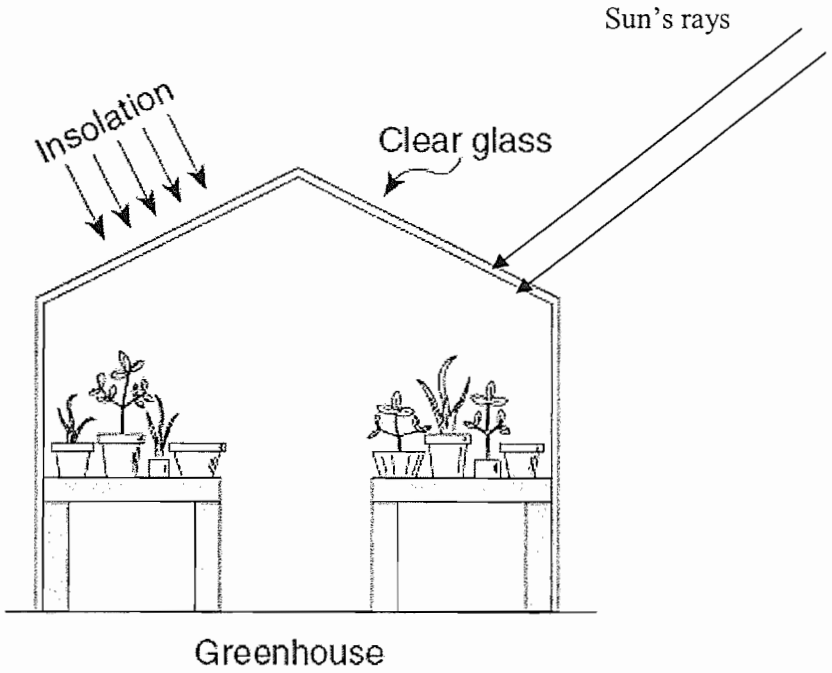
88. Hurricanes are different than tornadoes because

A) damage is more localized.
B) they are thin, wispy clouds made up of ice crystals.
C) they develop over water.
D) they are billowing and rounded in shape.

89. Which weather change is most likely indicated by rapidly falling air pressure?

A) Humidity is decreasing.
B) Temperature is increasing.
C) Skies are clearing.
D) A storm is approaching.

The diagram below shows a greenhouse.

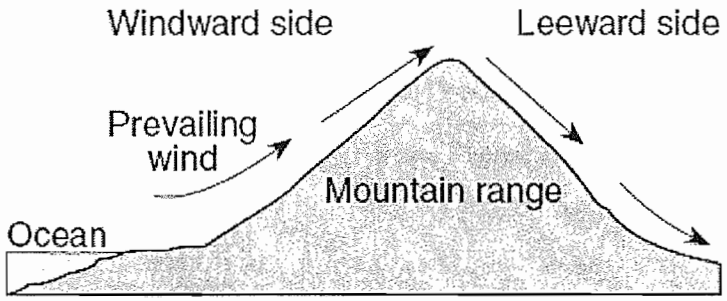


Greenhouse

90. What is the primary function of a greenhouse that would illustrate the "greenhouse effect"?

A) The glass allows all wavelengths of radiation in to enter and all wavelengths of radiation to escape.
B) The glass is used to trap all incoming solar radiation.
C) The glass allows short wavelengths of radiation to enter, but reduces the amount of long wavelength radiation that escapes.
D) The glass allows long wavelengths of radiation to enter, but reduces the amount of short wavelength radiation that escapes.

The cross section below shows the prevailing winds that cause different climates on the windward and leeward sides of this mountain range.



91. Compared to the climate conditions on the leeward side of this mountain range, the conditions on the windward side are usually

A) cooler and wetter.
B) cooler and drier.
C) warmer and wetter.
D) warmer and drier.

92. Which method is primarily responsible for energy transfer in the atmosphere?

A) conduction
B) solidification
C) convection
D) radiation

93. The greenhouse gas that contributes most to an increase in global warming is?

A) carbon monoxide
B) carbon dioxide
C) nitrous oxide
D) sulphur dioxide

94. Which of the following has the least influence on global weather patterns?

A) The equator is hotter than the poles.
B) The Earth's rotation.
C) Warm air rises.
D) Cold air can hold less moisture than warm air.

95. During El Nino

A) the surface temperature of the Pacific Ocean increases, the trade winds travel eastward, and rainfall increases along the coast of Peru.
B) the surface temperature of the Pacific Ocean increases, the trade winds travel westward, and rainfall increases along the coast of Peru.
C) the surface temperature of the Pacific Ocean decreases, the trade winds travel westward, and rainfall increases along the coast of Peru.
D) the surface temperature of the Pacific Ocean increases, the trade winds travel eastward, and rainfall decreases along the coast of Peru.

96. Ocean and wind currents are caused by

A) convection currents and the earth's rotation.
B) the force of gravity and landmasses.
C) the collision of warm moist air masses with polar air masses.
D) the collision of warm dry air masses from the south and moisture-filled Northern air masses.

97. The influence of Earth's rotation on air, or any object moving on Earth's surface is called the

A) Coriolis effect.
B) prevailing winds.
C) rotation effect.
D) spherical effect.

98. Ribbons of extremely fast moving air near the top of the troposphere are called

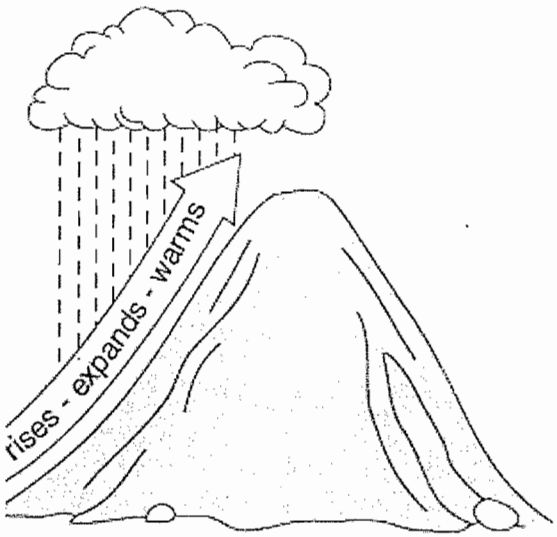
A) latitude air.
B) power streams.
C) prevailing westerlies.
D) jet streams.

99. Which direction do the prevailing winds blow in Manitoba?

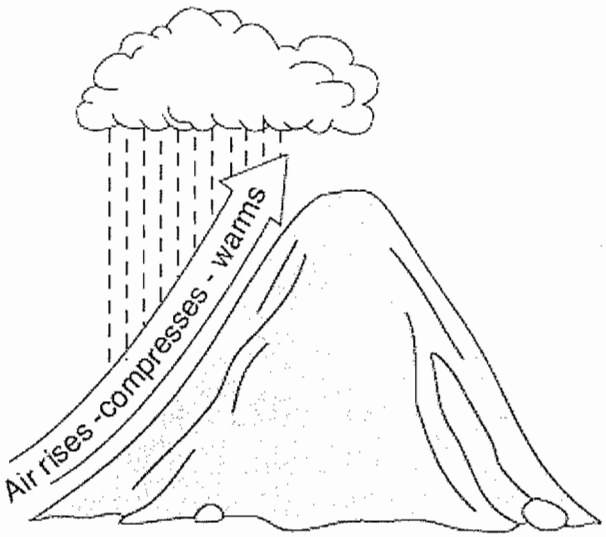
A) north
B) south
C) east
D) west

Go to next page

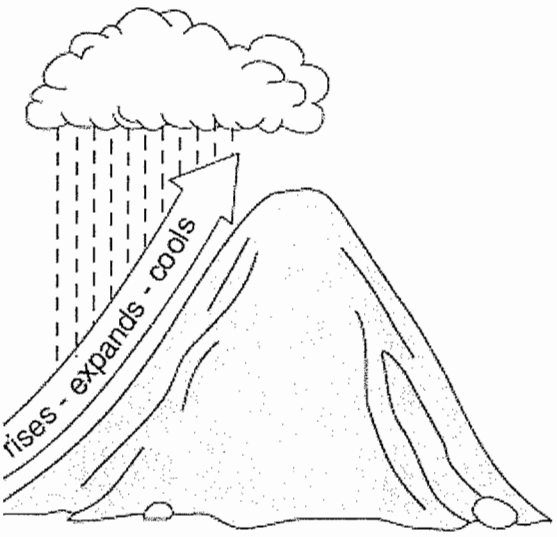
100. Which diagram <u>best</u> illustrates how air rising over a mountain produces precipitation?



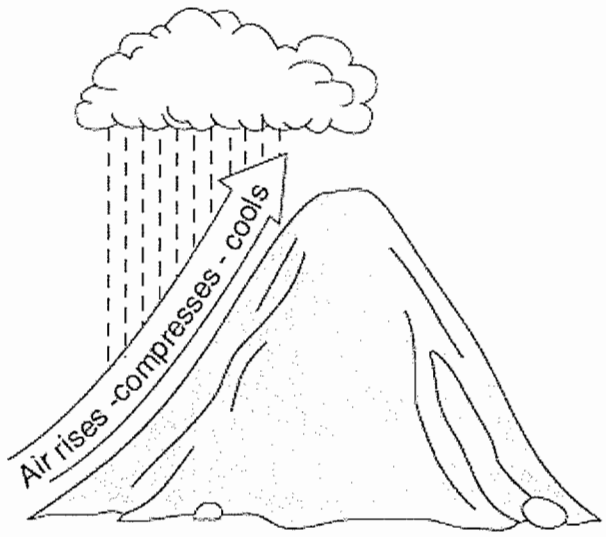
(1)



(3)



(2)



(4)

A) 1
B) 2
C) 3
D) 4

Appendix C

SC20F 2006 Exam Template

Weighting

Multiple Choice = 65% of exam grade

Extended Answers = 35% of exam grade

Each Cluster is given equal weighting in both multiple choice and extended answer.

Multiple Choice Section

Cluster 1: Dynamics of Ecosystems (19 MC Questions)

Outcome	Exam	Outcome	Exam	Outcome	Exam
1-01	Yes (3x)	1-02		1-03	Yes (2x)
1-04	Yes (4x)	1-05	Yes (3x)	1-06	
1-07		1-08		1-09	Yes (3x)
1-10	Yes (4x)				

Cluster 2: Chemistry in Action (19 MC Questions)

Outcome	Exam	Outcome	Exam	Outcome	Exam
2-01	Yes (1x)	2-02	Yes (5x)	2-03	Yes (1x)
2-04		2-05	Yes (2x)	2-06	Yes (2x)
2-07	Yes (1x)	2-08	Yes (4x)	2-09	
2-10	Yes (2x)	2-11	Yes (1x)	2-12	

Cluster 3: In Motion (19 MC Questions)

Outcome	Exam	Outcome	Exam	Outcome	Exam
3-01	Yes (5x)	3-02	Yes (4x)	3-03	
3-04	Yes (1x)	3-05		3-06	Yes (1x)
3-07	Yes (2x)	3-08	Yes (1x)	3-09	Yes (1x)
3-10		3-11	Yes (3x)	3-12	Yes (1x)
3-13					

Cluster 4: Weather Dynamics (18 MC Questions)

Outcome	Exam	Outcome	Exam	Outcome	Exam
4-01	Yes (2x)	4-02	Yes (2x)	4-03	Yes (7x)
4-04	Yes (5x)	4-05	Yes (2x)	4-06	
4-07		4-08			

Appendix D

AFFIDAVIT

I hereby affirm that all applicable procedures respecting the administration

of this Division examination, outlined in Operational Procedure 13005, have been followed.

Examination _____

School _____

Date & Time of Exam _____

Principal or designate's signature _____

Return to Assessment & Evaluation Coordinator
PSDC

Appendix E

2007-2008 SENIOR YEARS ASSESSMENT SCHEDULE

Oct.26/07

NOV. 2007 Gr. 12 Consumer Math Standards Test Project, Nov. 19-26; Portfolio due prior to Jan. 21

SEMESTER ONE

JAN. 2008	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
	Dec 31	1	2	3	4
	7	8	9	10	11
		Gr. 12 English Language Arts Standards Test, Jan. 8-11			
	14	15	16 MA20SA Spdsheet Project	17 MA30SA Project	18
	21	22	23	24	25
		EN20F, EN30S Jan. 21-25			
	Gr. 12 Consumer Math Standards Test	Gr. 12 Applied Math Standards Test, Jan. 22-23		Gr. 12 Pre-Calculus Math Standards Test	
Assessment Week: Jan. 28-Feb.1	28 a.m.: HI30S, PH40S p.m.: MA20SP/F, MA30SP/F	29 a.m.: MA20SA, PH30S, BI40S/F p.m.: CH30S	30 a.m.: MA10F/F, SC20F/F, BI30S/F	31 a.m.: MA30SA	Feb. 1 School Based

APR. 2008 Gr. 12 Consumer Math Standards Test Project, April 7-14; Portfolio due prior to June 10

SEMESTER TWO

MAY 2008	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
				May 1	2
	5	6	7	8	9
	12	13	14	15	16
	19 *Victoria Day*	20	21	22	23
	26	27	28 MA20SA Spdsheet Project	29 MA30SA Project	30
	Gr. 12 French Language Arts Standards Test, May 26-29				

JUNE 2008	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
	2	3	4	5	6
	Gr. 12 English Language Arts Standards Test, June 2-5				
	9	10	11	12	13
	EN10F, 20F, 30S June 9-13				
	Gr. 12 Pre-Calculus Math Standards Test	Gr. 12 Consumer Math Standards Test	Gr. 12 Applied Math Standards Test, June 11-12		
Assessment Week: June 16-20	16 a.m.: MA10F/F, HI30S, BI40S/F p.m.: MA20SA, MA30SA	17 a.m.: MA20SP/F, MA30SP/F, PH40S p.m.: BI30S/F	18 a.m.: SC20F/F, CH30S	19 a.m.: PH30S	20 School Based
	23	24	25	26	27