INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

THE UNIVERSITY OF ALBERTA

A VIEW INTO PERFORMANCE ASSESSMENT IN SCIENCE

BY

PAUL WOZNY 

A DISSERTATION SUBMITTED TO THE FACULTY OF

GRADUATE STUDIES AND RESEARCH IN PARTIAL

FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

DOCTOR OF EDUCATION

DEPARTMENT OF SECONDARY EDUCATION

EDMONTON, ALBERTA

SPRING, 1998

The University of Alberta

Faculty of Graduate Studies and Research

The undersigned attest that they have read and recommend to the Faculty of Graduate Studies and Research for acceptance, the dissertation "A View Into Performance Assessment in Science" submitted by Paul Wozny in partial fulfillment of the requirements for the degree of Doctor of Education.

.....
(Advisor)

.....

.....

.....

.....

.....
(External Advisor)

Date: April 17/1998

Abstract

This collaborative research project involved the design, implementation, and analysis of five sets of performance assessment activities with a group of eight science teachers at an urban composite high school in Alberta.

These science teachers shared their initial impressions of the performance assessment process and the feasibility of this mode of assessment in their classrooms. The teachers were somewhat reserved in the feasibility of enacting performance assessment tasks, largely due to the time constraints associated with larger classes.

The five sets of performance assessment tasks designed and implemented by the research group included: 1) basic electronics (science nine), 2) density problems (science nine), 3) microscope skills (science ten), 4) uniform motion (science ten), 5) acid/base identification and neutralization (science ten). Analysis of the performance assessment results included standard deviation, Pearson's Product Moment Correlation Coefficient, and face validity evaluation. Inter-rater reliability varied from 0.83 to 0.91 (Pearson's Product Moment

Correlation Coefficient) over the entire group of performance assessment tasks, which indicates very strong inter-rater reliability. These results reinforce the research of Gipps (1994) which found that the inclusion of "clear rubrics and training for markers, and exemplars of performance at each point or grade, levels of IRR (inter-judge reliabilities) can be high" (Gipps, 1994, p. 104). The face validity of the performance assessment tasks was also seen as very strong due to the close fit with suggested activities in the science curriculum.

The participating teachers shared a strong appreciation and approval of the performance assessment process in the science classroom after designing and implementing the five sets of performance assessments, but had some reservations about the time involved in the set-up and implementation.

In the appendix of this dissertation, I included a teacher's handbook that provides a simple and quick overview of the performance assessment process coupled with examples and lessons learned.

Acknowledgements

First off, I would like to thank my wife, Addy and son, Scott. The love shared in our lives allows for the growth and freedom of the learning process. I would also like to thank my advisor, Dr. Wytze Brouwer, for his patience and knowledgeable contributions to my educational experience. The inspiration and love of writing shared by Dr. Jim Parsons also motivated me to continue writing as a life-long hobby.

Equal partners in this research project are the fabulous science staff at Beaumont Composite High School. Their collective enthusiasm and love of science allows my passion for research, science and technology to flourish!

Table of Contents

List of Tables

Chapter One - Introduction

Motivating the Researcher

This research project began as a plan to share real-world experiences of secondary education science teachers attempting performance assessment tasks in a collaborative fashion with students in a secondary (grades 9 to 12) education environment. As an instructor of both science education and technology education for the past sixteen years, I have been often challenged by the integration of performance assessment tasks in my daily pedagogical routine. The time requirements in setting up performance assessment tasks coupled with the challenges of dealing with a large number of students in the classroom could potentially "turn off" a teacher toward performance assessment tasks as a mode of assessment. In my particular case, I was compelled to pursue this line of research to increase my understanding and effectiveness in utilizing performance assessment tasks in both my science and technology education classes.

The aspects of validity and reliability of the performance assessment process have also caused much consternation in my integration of this "hands-on" mode of assessment. How can a teacher be certain that the particular performance assessment task they have implemented is indeed testing what he/she thinks it is testing? How consistent are the teachers' appraisals of students, from one individual

to the next? Herein lies the basis of this research project which involved working with a group of eight secondary science educators, at an urban composite high school, over the period of one school year (September, 1995 to June, 1996). This research group worked collectively to complete a variety of performance assessment tasks from grade nine to eleven levels over the subject domains of physics, chemistry, and biology.

Research for my Master's degree was focused in the domain of computer-based instruction and testing coupled with hands-on experimentation emphasizing physics and advanced technology. The end-product of this research included the fabrication of discrete instructional modules packaged in a small, transportable suitcase. Each module included both computer-based instruction and testing along with the actual hardware necessary to complete the hands-on portions of the particular instructional module. One module in the domain of Lasers included the student attempting a hands-on lab activity with a low-powered (1 milliwatt) helium neon Laser. The student was challenged to guess the distance of a Laser from the side of a building in which the Laser was focused on. The student also completed a detailed write-up of the results of his/her research into the rate of Laser beam expansion over specific distances. He/she then extrapolated the probable distance that the Laser was from the side of this particular building. It was the evaluation of the student's actual manipulation of the Laser (the

experimental apparatus) and his/her abilities to accurately measure, observe, analyze, draw conclusions, and express sources of error which motivated me to continue my research at the doctoral level in the domain of performance assessment.

A recent curriculum change in technical and vocational education from Alberta Education has introduced a new set of secondary education courses known as "Career and Technology Studies." This set of courses utilizes specific learner expectations coupled with competency based assessments according to set criterion. This criterion-based performance assessment methodology has placed many science and technology teachers in the challenging position of having to create and enact performance assessment tasks in their particular pedagogical situation. Moreover, Alberta Education has also distributed sample performance assessment tasks in general science, physics, chemistry, and biology. These sample tests are undoubtedly a sign of the future trend towards integration of performance assessments in the overall evaluation schemata of the student. This places the teacher in the dynamic situation of having to "get to speed" about the performance assessment process and hopefully pursue this mode of assessment in a reliable and valid manner.

I had the opportunity to meet with my own high school physic's instructor shortly before I had decided to continue my graduate study.

When I visited his Physics 30 class, he was in the process of completing a field study of a performance assessment task designed by Alberta Education. While sharing old memories with my former instructor, I became quite fascinated with the students who were in the process of completing a set of performance assessment tasks. Each student had been given a number of cups with differing numbers of pennies distributed in each cup. The challenge for each student was to determine what the fundamental unit (the mass of a single penny) was with respect to the group of cups. This performance assessment was an obvious attempt to replicate the challenges of determining the fundamental charge of an electron, which was addressed by Millikan in his now famous oil drop experiments. What I found quite interesting about this performance assessment task was the fact that each student appeared to be quite determined to solve this problem which involved real apparatus in a hands-on experimental context. Not only had they attempted to solve this dilemma, but they also created and followed their particular experimental theories and methods.

My subsequent conversation with my former physics teacher revealed many issues which reinforced my determination to better understand the dynamics of the performance assessment process. First, he shared his difficulty in implementing this performance assessment task which involved each of his 30 students having their own, individual

work-station which did not allow them to see the fruits of each other's work. This problem was solved by using a computer education facility, which had partitioned walls between each computer thus minimizing attempts to see what others were doing. Second, the teacher had to invest many hours of preparation for this activity in terms of materials and apparatus, which is no small task in large classes. Third, the instructions which were sent with this field test performance assessment were quite minimal and caused much anxiety and uncertainty with the teacher (physics 30 instructor) who was trying to implement it. Fourth, the scoring rubrics did not provide examples to clarify to the teacher exactly how to grade each of the students both while performing the experiment and assessing the actual write-up of their particular strategies and results. Fifth, some Physics 30 classes had exposure to this performance assessment field test at an earlier date, and may have shared this exposure with students in the other classes. As a result, these students knew what the performance assessment entailed. This prior knowledge could reduce the fidelity of this task by allowing students time to discuss and solve this problem before even stepping into the classroom. Is it possible that students with friends at a different school that had implemented these performance assessment tasks at an earlier date could have been informed of the nature and content of the particular performance tasks? As one can appreciate, the physics

teacher found this process challenging, time consuming, and perplexing; yet, through the entire process he felt that this mode of assessment was a necessary ingredient in the overall evaluation of the student.

Armed with the actual experience of viewing students attempting performance assessment tasks in the physics classroom, coupled with the sharing of impressions with the teacher directly affected by this process, I felt quite motivated to pursue this line of research. The problems which I had witnessed (time, reliability, validity, fidelity) would form the nucleus of this research project.

Chapter 2

Issues and Problems on Performance Assessment in Science

In my analysis of issues and problems in performance assessment, I drew from four research projects: 1) Research on Assessment in Science (1994) by R. Doran, F. Lawrenz, and S. Helgeson. 2) Laboratory Performance Tests for General Physics (1959) by G. Kruglak, and C. Wall. 3) Beyond Testing (1994) by C. Gipps. 4) Evaluation of Science Practical Activity (1980) by G. Jeske. The challenge and importance of including performance assessments in the science classroom is certainly not new, as witnessed at the Cavendish Laboratory in Cambridge University during 1887. Performance tests were given to physics students and yielded some very perplexing findings (Kruglak and Wall, 1959). What amazed the examiners in this pioneering study at the Cavendish Laboratory in 1887, was the large gap that existed between the abilities of the physics students in written exams as compared with performance examinations. This finding is shared by R. Wilberforce in his paper A History of the Cavendish Laboratory 1871 - 1910:

> "There were two papers in Elementary Physics, including mechanics, but no oral or practical examination in these subjects was held, and doubts began to be felt as to the adequacy of this test and the value of the knowledge likely to be acquired by candidates who were preparing for it. About the year 1887 these doubts were set permanently at rest by the institution of a simple oral examination in physical apparatus . . . . we heard for the first time of the man who recognized in a thermometer a machine for determining the

> specific gravity of water, although, when he was furnished with the thermometer, a basin of water, and a bit of string, he failed to achieve any numerical result. Again there was the man who said that a compass needle mounted over a graduated circle was an instrument for determining latitude and longitude. "What," said the examiner aghast, "can you determine the latitude and longitude with this?" "No sir" said the man, "but you can sir," a trustfulness which deserved a higher reward than it received . . . . Very soon after the year 1887 the University of London instituted a simple practical examination in addition to the papers in Physics hitherto set in the Preliminary Scientific Examination in Science for the degree B.Sc.. The early examiners, of whom I was one, were faced by a mass of absurdity even greater than that which had been revealed at Cambridge, and the most disquieting part of the matter was that very often candidates who seemed from the written work to have a real knowledge of their subject made the grossest possible blunders when confronted with the actual apparatus of which they were able to furnish perfect diagrams and descriptions on paper." (Kruglak and Wall, 1959, p. 13 - 15)

Even back in the latter part of the 19th century, it was realized that although one may be able to display great capabilities on a written exam, this did not necessarily correlate to excellence in the actual utility of one's abilities in performance based assessments. The previously mentioned example of the implementation of performance assessments in physics at the Cavendish Laboratory in 1887, indicated that the *abilities of the student to transfer his/her knowledge to the solution of problems with real equipment in a hands-on environment did not seem to correlate well with theoretical written exam knowledge* (Kruglak and Wall, 1959, p. 13 – 15).

The ramifications of this early ground-breaking research are vast,

and perhaps only recently appreciated by educators and administrators. It was this type of research that motivated me to develop, design, and implement the advanced technology programs at Beaumont Composite High School in Beaumont, Alberta. My intents were to synthesize the math, science and technology education fundamentals in the hope of creating hands-on methods of learning that reinforced the dynamic interplay between mathematics, science and technology. By this process I did not simply intend that all students' activities should be tied to hands-on performance tasks, but rather an integrated approach of learning which included the hands-on aspect of scientific investigation and problem-solving, as well as the traditional paper and pencil theoretical concerns. This integrated approach is similar to many physics, chemistry and biology learning situations in post-secondary institutions where the learner is actively engaged in hands-on experimentation as well as classroom lecture.

An issue that has dominated the field of performance assessment is that of reliability. Reliability concerns the repeatability and consistency of the measurement process. Doran et al. (1994) describe the issue of reliability in performance assessment as involving various aspects of reliability such as "within one test (i.e., internal consistency), across time (i.e., stability), across form (i.e., equivalency), and across raters (inter-rater agreement and correlation)" (Doran, Lawrenz, and

Helgeson, 1994, p. 425). Tamir (1974) researched reliability in performance assessment in the domain of biology laboratory exams and, in particular, focused in the domain of inter-rater reliabilities which is also the domain of research which I have pursued in continuation of the pioneering work of Tamir:

> "Tamir reported that when the individual raw scores in each of the six skills was used, the Cronbach reliability value ranged between .56 and .69 with a mean of .67. The inter-rater agreement was obtained with six examiners, in 1971 and 10 examiners in 1972. Each test was scored by two examiners, who were assigned to the tests at random. Each examiner scored between 30 and 100 papers. The correlation between the total raw scores given by each pair of examiners were computed. These values were regarded as measures of inter-rater reliability coefficients. The values ranged between .54 and .89 in 1971 and .57 and .87 in 1972. The mean inter-rater correlations were .82 and .79 in 1971 and 1972 respectively."
> (Doran, Lawrenz, and Helgeson, 1994, p. 425)

This research by Tamir led to his working with other researchers (Nussinovitz and Friedler) in formulating a guide for the assessment of abilities in performance exams. This guide was utilized by Hargraves and Lynch (1988) in the implementation of performance assessment tasks in Tasmania. They (Hargraves and Lynch) concluded that "teacher confidence in the reliability and validity of the marking process was considerably enhanced after using the inventory" (Tamir's Practical Test Assessment Inventory) (Doran, Lawrenz, and Helgeson, 1994, p. 425). Tamir's research points to the importance of very clearly delineated

scoring rubrics coupled with appropriately trained evaluators in the improvement of reliability in performance assessment.

I have a keen interest in the inter-rater reliability of performance assessment in science and technical education. Performance assessment is certainly not new to me. Back in 1974, when I began my initial training as a pilot, I was evaluated both on written examinations from the Ministry of Transportation (Aviation) and performance examinations, which involved the actual flying of an aircraft through a variety of situations and inflight problems and emergencies. Obviously the consequences of improperly trained pilots are very severe, and the integration of performance assessment tasks has been a cornerstone of the pilot training process. It has long been recognized in the world-wide aviation community that theoretical knowledge in aviation does not necessarily transfer to inflight skills and emergency handling capabilities.

If one is tested extensively in performance, then the instruction process will most likely respond to this reality (Doran et al., 1994). I suspect that physics teachers would likely spend more time in hands-on laboratory activities if the diploma exams incorporated a performance assessment component that counted for a significant portion of the student's grade. This could raise the concern that teachers may teach to the exam on performance skills as they might for traditional multiple-

choice exams. One could also argue that teachers should teach certain tested performance assessment tasks, as in the case of professional pilots whose parameters of performance are well known before actual flight tests. Undoubtedly the inclusion of performance exams would enhance the evaluation process, as was evidenced by the previously described experiences at the Cavendish Laboratory of Physics in 1887.

Concerns over the implementation of performance assessment tasks have been shared by The New York State Education Department in 1992 and include that it (performance assessment) may:

> 1. be untried and may not prove to better measurement of student ability;
> 2. be considerably higher in cost than standardized testing;
> 3. be identified by some psychometric experts as less reliable and less valid than traditional tests;
> 4. be just as vulnerable to being "taught to" as multiple choice style testing;
> 5. be liable not to result in the better performance by those who fare poorly on the standardized tests since performance testing reflects reality;
> 6. be likely to reveal wider gaps in the achievement of disadvantaged students since changing the type of assessment will not change outcomes; and
> 7. necessitate that teachers will be required to teach in new ways consequently requiring professional development and re-education.
> (Doran, Lawrenz, and Helgeson, 1994, p. 426)

Is it possible that these concerns are also experienced by the decision makers at Alberta Education or by some of the larger school boards in Alberta? I suspect that cost is a large factor in any adoption of change in the public education process. Performance assessment does require

additional time and effort by teachers and administrators. In the present political climate of Alberta, with funding towards education being drastically reduced, does it seem realistic to expect facilitation and support in the implementation of performance assessment? Gipp's (1994) search into utility of performance assessment examinations in public schools allays some of the concerns shared by the New York State Education Department in 1992: "clear rubrics and training for markers, and exemplars of performance at each point or grade, levels of IRR (inter-judge reliabilities can be high" (Gipps, 1994, p. 104).

In Laboratory Performance Tests for General Physics (1959) by Kruglak, G. and Wall, C., a critical review of published research in the domain of laboratory instruction was provided. The following criteria were utilized by Kruglak and Wall in their analysis:

> 1. Did the investigator make a sufficiently detailed report of his work so that it can be properly appraised?
> 2. Did the investigator use the best contemporary experimental procedures and techniques?
> 3. Were the research data subjected to the best analytical treatment known at the time of the study?
> 4. Has the investigator been reasonably cautious in interpreting the data and drawing conclusions?
> 5. Did the investigation make a contribution to the field?
> (Kruglak and Wall, 1959, p. 136 - 137)

A colloquium paper by Gerhard Jeske (1980) Evaluation of Science Practical Activity submitted toward completion of the degree of Master of Education in the Department of Secondary Education at the University of

Alberta consisted of a detailed survey of literature on the topic of evaluation of practical activities in science. Jeske provided a rich and holistic qualitative appreciation on the issue of practical evaluation in science. Jeske (1980) included a rationale for practical assessment based on the pedagogical perspectives of Gagne, Lehman and Blees. This rational stated:

> "Feedback during practical activities appears to be a necessity. Without evaluation to provide feedback it would seem to be impossible for the students to adequately determine how well they are meeting the desired performances, or to know when they have misconceptions about the principles, scientific processes, manipulative skills, attitudes, values, etc. being learned. Thus, evaluation is essential in helping students to guide their learning, or in helping them to get their bearings." (Jeske, 1980, p. 20)

Jeske further illustrated a conceptual framework for practical assessment based on Klopfer's Model, Bloom's Taxonomy, Nay-Crocker Affective Inventory, Nay's Process Inventory, and Simpson's Classification System of Psychomotor Skills. Jeske contends that the need for performance assessment in the high school science curriculum is high. He based his conclusions from an evaluation of methods suitability in the domains of: 1) laboratory practical exams, 2) paper and pencil tests, 3) oral exams, 4) multimedia format, 5) laboratory reports, 6) student self-evaluation (Jeske, 1980, p. 141). This process led him to the conclusions that:

> If practical activities are to maintain a central position in the science curriculum then it is essential that extensive evaluation of these activities occur since, it is generally accepted that objectives not reflected in evaluation procedures are usually neglected by students . . . . practical activities need to include elements from the content, process, psychomotor, and effective areas . . . . practical activity should not concentrate exclusively on assigning of grades or scores to students, but should also emphasize diagnosis. Such assessment can be used to improve both student progress, and the quality of instruction being offered . . . . a sufficient number of practical activity evaluation methods are available; however most of these are not being used extensively enough in student evaluation . . . . testing is predominately on the pencil-and-paper mode, and this indicates deficiencies in practical activity instruction, since certain facets of practical work cannot be evaluated effectively this way . . . behavioral objectives being developed for practical activities are generally inadequate with respect to the affective and psychomotor areas . . . . Diagnostic evaluation of practical activity needs emphasizing since it can serve to indicate student weaknesses, instructional deficiencies, inadequate behavioral objectives, and the presence of ineffective evaluation instruments . . . . Descriptions in the literature are scanty with respect to the criteria needed to select suitable practical activity evaluation instruments, and likewise for the standards applied in assessing the quality of performance for specific evaluation devices.
> (Jeske, 1980. p. 139 - 142)

An informal survey of the science staff at my particular composite high school illustrated a strong research interest into the feasibility, reliability, and validity of the performance assessment process. These issues brought up by the staff are consistent with the research topics of Doran, Lawrenz, and Helgeson (1994), Tamir (1974), and Hargraves and Lynch (1988) in the domain of inter-rater reliability and validity of the

performance assessment process in the secondary education science classroom.

Chapter 3

A Selected Review of Performance Assessment in Science Education

When one investigates the past and present of performance assessment, you have embarked in a diversified journey of assessment paradigms which although quite varied, do have some similar characteristics which appear with repeatable cadence. These similar characteristics include the fundamental aspects of assessing actual psychomotor (hands-on) manipulation of certain variables and/or problem solving within the eyeshot of a moderator. The definition shared by Stiggins and Bridgeford (1982) states:

> "Performance assessment is defined as a systematic attempt to measure a learner's ability to use previously acquired knowledge in solving novel problems or completing specific tasks. In performance assessment, real life or simulated assessment exercises are used to elicit original responses which are directly observed and rated by a qualified judge." (Stiggins and Bridgeford, 1982, p. 1)

Kruglak and Wall (1959) share a similar perception of performance assessment: "In a laboratory performance test the student must deal directly with real apparatus and with an actual physical situation in order to succeed. He must not be able to circumvent the apparatus." (Kruglak and Wall, 1959, p. 12). Performance assessment in science is not a recent innovation. Performance testing in physics at the Cavendish Laboratory at Cambridge University was implemented in 1887 (Kruglak and Wall, 1959, p. 13)! These performance tests provided a

means of examining a student's practical skills and comparing his/her achievement with that in written examinations. As mentioned in the previous chapter, it was interesting to note that students did not necessarily succeed in their performance tests with the same degree of achievement as in their written exams. In fact, many students demonstrated what the examiners considered "abysmal failures when confronted with a real piece of apparatus" (Kruglak and Wall, 1959, p. 15).

Boykoff Baron (1991) holds a view of performance assessment where it:

> "becomes the culminating activity of a unit and provides opportunities for students to synthesize their knowledge, make connections, and deepen their understanding of major concepts, . . . creates situations that are intended to foster the development of deeper levels of understanding, . . . blurs the edges among assessment, curriculum, and instruction."

She (Boykoff Baron) relates that the methods of assessment in today's "era of accountability" drive the methods of instruction that occur in many classrooms. She concludes that performance testing should be a necessary addition to multiple-choice tests to foster instruction that will demand much more than memorizing information and the danger of teaching to the exam. Is it indeed possible that a Physics 30 teacher might skip certain laboratory activities to maximize time spent on the theoretical material covered on the province-wide diploma exam? Could

pressure for high achievement on diploma exams by a school's administration, which receives more funding for higher scores, possibly encourage "teaching to the exam?" If the diploma exam does not include a performance assessment aspect, will teachers cover the laboratory experiments with the same effort and rigor as the theoretical constructs of the course? One cannot help but appreciate that the methods of assessment and the consequences of their process can affect the methods of instruction!

Welford (1990, p. 53) contends that a "worthwhile aim of practical assessment ought to be to ensure that it takes place in the context of purposeful practical work." Welford's perceptions on practical assessment appear grounded not only in the manipulation of apparatus but also in the purposeful synthesis of practical skills in a holistic problem-solving capacity. Welford (1990, p. 37) contends that current trends in the assessment of practical work in science classrooms are undergoing a shift from:

> "an illustrative, confirmatory or discovery function - - with its apparent afterthought of increasing students' interest and motivation - - to one which additionally seeks to develop and rehearse investigative skills. Developments in assessment are now offering criteria against which to judge abilities such as handling the variables of the experiment, deploying the strategies of investigation and using the techniques of measurement, observation, data organization, interpretation and deduction."

This view of practical assessment emphasizes the abilities of the

assessment procedure to not only increase student motivation and practical psychomotor skills in the manipulation of apparatus but also examine investigative constructs of the particular participant. One could appreciate this view of practical assessment as being consistent with the internal schemata of a constructivist. The students' assumptions of the investigative process could be evidenced in a well-scripted performance examination. I find this prospect interesting from the diagnostic perspective, in that an instructor could have an opportunity to gain a deeper appreciation of the learner's constructs of the investigative process.

Tamir (1985) shares three different formats for individual performance examinations. The first type involves the examiner observing unobtrusively while the student performs various required tasks. The examiner then grades these tasks according to a preset scoring checklist or rubric. The second mode of performance assessment parallels the clinical interview process utilized by Piaget (Doran, Lawrenz, and Helgeson, 1994). In this mode the examiner questions and probes the examinee in the attempt to understand his/her perceptions. Tamir's description of a third mode of performance assessment entails the examinee performing an oral exam on a particular topic "based on concrete phenomena or materials " (Doran, Lawrenz, and Helgeson, 1994). Such examinations often follow experiments or projects.

Fitzpatrick and Morrison share two thoughts on performance testing:

> "1. Performance testing implies the simulation of an actual situation.
> 2. The degree of realism (fidelity of simulation) can range from the complete artificiality to the observation of the task as it is performed in the real world (eg., job sample test). At one extreme, there are tests whose fidelity of simulation ends at the title of the test; at the other extreme, there are tests that are very realistic but not very practical or useful. To some extent, any standardized measure of performance is less than truly authentic because the process of observing the performance invariably influences the behavior of the performer." (Finch, 1991, p. 4)

The second point shared by Fitzpatrick and Morrison quite clearly delineates the inherent aspect of performance as being somewhat affected by the observing process. This would lead one to appreciate the dynamic interplay of the observation process and the scoring rubrics in performance assessment.

Britain has a long history of integrating performance assessment activities into the examination structure of both secondary and post-secondary students. Stephen Knutton (1994) shared the advantages offered by the integration of performance assessment in science within the British education system:

1) the elimination of chance failure in a one off situation
2) providing a richer and more varied experience of practical work
3) enabling a wider range of skills to be assessed (such as attitudes)
4) greater reliability (teachers are in the best position to assess

students' practical skills because they see them over an extended period of time.)
5) permitting theory and practice to be more closely linked
6) becoming an integral part of the teaching and learning process (formative rather than just summative)
(Knutton, 1994, p. 155)

The United Kingdom witnessed the introduction of the "General Certificate of Secondary Education" examinations in 1988, which required teachers to make assessments about the practical skills of their students in science (Knutton, 1994). Initial problems with the practical assessments focused upon the:

> "shortages of apparatus for suitable experimental work and a dearth of well tried out experimental activities capable of yielding the necessary assessments of skill. Within a short time the ingenuity of science teachers produced a wealth of published resource materials. Problems still remained on the management aspects of making valid assessments of groups of up to thirty students. These logistical problems were a real concern for many teachers who were, not surprisingly, daunted at the thought of trying to assess the practical skills of such a large group on the same occasion. As time has progressed a range of strategies for coping with these situations has been devised . . . . These include:
> 1) devising exercises that leave a record
> 2) making use of an additional assessor
> 3) routinely checking during normal practical work
> 4) adopting a "stations" approach
> 5) making use of teacher demonstrations
> 6) using questions in written examinations"
> (Knutton, 1994, p. 156)

Reliability in Performance Assessment

Reliability is often associated with a test's "capacity to give similar results for a particular student when he/she is tested on different

occasions." (Ashworth, 1982, p. 106) A word often associated with reliability is *consistency*. Will this test I administer to my students give consistent scores when conducted several times? The challenge which performance assessment faces is based in the dynamic latitude which exists in the evaluative strategies coupled with the necessity for a repeatable "ruler" with which to measure an examinee's abilities. One must appreciate that evaluation of performance is extremely challenging in comparison to pencil and paper exams, in that the interpretation of the examinee's performance is subject to observational bias either intentional or unintentional by the moderator/s.

An example of the challenge of judging performance assessment is exemplified in Olympic figure skating events where a group of judges adjudicate the performance of each skater. This issue was certainly a controversial one during the pairs ice skating event in Nagano, Japan during the 1998 Winter Olympics. What one witnessed was the wide variation and range of scoring for the various contestants from judge to judge.

In a multiple choice math or physics examination there may be only one correct response for each question, yet in the process of judging a student's performance on a mental and psychomotor (hands-on) activity, one may be faced with a multitude of variables which interface with the actual performance. It is in the appreciation of the multi-

variable nature of performance assessment that one faces the challenges of reliability that enter in the moderation (judging) process (Brigance and Hargis, 1993). This theme of consistency is the central focus of reliability:

> "The degree to which a test can be relied on to give consistent results when administered to the same or similar groups is referred to as its reliability. The reliability of a test is the degree or extent to which the results of the test are stable and can be trusted."
> (Brigance and Hargis, 1993, p. 64)

Reliability can be further delineated into test-retest reliability and inter-rater reliability. Test-retest reliability corroborates that a particular performance assessment will bear consistent results when conducted over several occasions (Brigance and Hargis, 1993). One can research the test-retest reliability of a particular performance assessment by conducting it several times with the same individual. If the results do not vary to any critical degree, one can feel comfortable that test-retest reliability is conformed (Brigance and Hargis, 1993). In performance assessments one must appreciate that the possibilities of students' knowledge of the particular performance assessment may affect the retest scores. One might expect scores to improve as the students execute the same performance assessment at a later date. Test-retest reliability is crucial to a performance test's usefulness. If one is not able to demonstrate consistency in the scoring of a particular performance

assessment, then this exam may not be used as a reliable indicator of a student's abilities.

In the process of teaching my own science and technology education programs, I have found that the issue of consistency in performance assessments was very much tied to the mode of scoring rubrics which I had established for a particular performance assessment task. Scoring rubrics refers to the scoring structure and categories that are utilized in the evaluation of a particular performance assessment task.

The utility of effective scoring rubrics has demonstrated the capability of increasing my own test-retest consistency in the evaluation of student's projects. This was not always the case in my early teaching career. When I initially began teaching, I would evaluate a particular student's work in the industrial arts or science laboratory based on my own personal rating scale which unbeknown to me at the time, was a "rubber" ruler of sorts. After a few months of this mode of assessment I quickly realized that both the students and myself required a clearer appreciation of the various aspects of evaluation which would be incorporated into the assessment process. The reason for this change was due to my awareness of the relative inconsistency of the assessments of student's projects in the shop when viewed at a later date and somewhat surprise with the marks given. This was clearly brought

to my attention by a student who commented that he felt his electronic project was constructed with the same quality as another student, yet he received a lower grade upon evaluation. The difference in the two grades was quite small, yet it prompted me to utilize a structured scoring guide that broke down various aspects of the particular performance assessment task. My subsequent use of these scoring rubrics appeared to increase the consistency of my evaluations. Moreover, the students were also made aware of the specific aspects of their particular project that would be evaluated coupled with exemplars of the various levels of accomplishment.

Tamir, Nussinovitz and Friedler formulated a guide for the assessment of abilities in performance tasks which was designed to increase test-retest reliability. This guide was utilized by Hargraves and Lynch (1988) in the implementation of performance assessment in Tasmania. They (Hargraves and Lynch) concluded that "teacher confidence in the reliability and validity of the marking process was considerably enhanced after using the inventory" (Tamir's Practical Test Assessment Inventory) (Doran, Lawrenz, and Helgeson, 1994, p. 425). Tamir's research pointed to the importance of very clearly delineated scoring rubrics coupled with appropriately trained evaluators in the improvement of test-retest reliability in performance assessment.

Swezey (1981) suggested that test-retest reliability in performance

assessments should be checked by retesting examinees using the same performance task:

> "administer the test to the same examinees twice, relatively close together in time. Only about one day should elapse between the first and second administrations of a criterion-referenced test where the purpose is to establish test-retest reliability estimates."
> (Swezey, 1981, p. 145)

In the process of retesting individuals on a particular performance assessment task, one must ensure that the conditions and context of the process are consistent from first application to the second application. If extraneous learning/mastery occurs between the first and second test applications, one could expect changes in the abilities of the examinee with regard to a particular performance task. Moreover, if the examinee is aware of the fact that he/she will be retested on the same performance task, he/she might practice and/or study that particular task and increase the probabilities of successful completion.

> "A second important point in assessing test-retest reliability is that the subjects in the reliability sample cannot be informed that they will be retested. The retest must come as a surprise, otherwise the subjects may prepare for the retest during or after the test's initial administration. For purposes of reliability assessment, it is inappropriate for examinees to practice or to study the tasks or skills between the first and second test administrations. Nor is it appropriate for examinees to attempt to recall the test in detail. Test-retest reliability assessment presumes no practice between test administrations, and equivalent testing conditions on both administrations. The equivalent condition requirement applies not only to the specific testing conditions, but also to

extraneous and environmental conditions."
(Swezey, 1981, p. 145-146)

The examiner is challenged to ensure that the examinee and conditions are duplicated in exacting detail from first to second application of a particular performance assessment. Would one not expect the process of experiencing a particular performance assessment to initiate subtle changes in the cognitive appreciation of that particular task where upon the second application the examinee is not responding with the exact same mind-space and preconceived notions? This point most certainly provides challenges in establishing test-retest reliability within the realm of performance assessment.

Inter-rater Reliability

Inter-rater reliability is often viewed as a critical indicator of consistency in the performance assessment process. Unlike a multiple-choice exam where scoring requires little subjective thought from the scorer other than a good key and an accurate eye, performance assessments often demand significant amounts of subjective judgement from the moderator/s (Brigance and Hargis, 1993). In a situation where five examiners observe a student completing a performance task, it is possible that their scores may not agree with each other. If the scores are very close then one would appreciate the inter-rater consistency as good, but if the scores amongst the five examiners varied quite markedly,

then one would view this assessment situation as unreliable in terms of inter-rater consistency. Pioneering research on inter-rater reliability in performance assessment tasks was done in 1912 by Starch and Elliot:

> "A classical study or example is the research work reported by Starch and Elliot (1912) regarding the lack of reliability between English teachers scoring English papers. They had two compositions evaluated by 142 English teachers. They found a 47 point range of scores on a 100 point scale. One paper received 15 percent failing scores and 12 percent above 90."
> (Brigance and Hargis, 1993, p. 67)

In appreciation of the reality that inter-rater reliability is dependent on the various moderators/judges/examiners coming up with consistent scores, one cannot help but appreciate the importance of well-trained examiners coupled with clearly delineated scoring rubrics in the attempt to increase inter-rater reliability in performance assessments. Priestly (1982) shares three common errors amongst moderators that can decrease inter-rater reliability in the performance assessment process: 1) *personal bias*, 2) *halo effect*, 3) *logical error.*

Personal bias can cause a rater to score the examinees amongst a mean previously determined by that particular examiner. "Generous scorers tend to rate everyone high, severe scorers rate everyone low, and others lump everyone in the middle of the scale." (Priestly, 1982, p. 127) The challenge to moderators in performance assessments is to establish a common focal point that is appreciated with the same degree of

qualitative and quantitative rigor. The utility of scoring rubrics coupled with examiner training can reduce the effects of personal bias. The scoring rubrics must delineate a scoring procedure and standard common to all examiners.

The issue of inter-rater reliability was addressed in Educational Assessment by Brigance and Hargis (1993). Their collaborative investigation into practical assessment in the American school system resulted in:

> "the more judgments an examiner must make in scoring a test, the more precisely the scoring procedure should be structured. Strict adherence to these procedures will be required to assure the reliability of scoring. Many individually administered tests require a considerable amount of judgement in scoring or rating. The administration procedures and the scoring procedures are spelled out in detail. These procedures must be followed in order that the results can be relied on to represent the tests' purposes."
> (Brigance and Hargis, 1993, p. 68)

A situation known as "halo effect" occurs when a scorer's general impression of a person induces the scorer to rate the person the same on all dimensions or the same over a period of time on many dimensions." (Priestly, 1982, p. 128) The halo effect can occur when an examiner tends to rate an examinee in a positive manner throughout a particular performance assessment based on a few traits such as personal appearance, and/or personality (Priestly, 1982).

I personally can relate to the halo effect in my own evaluation

practice within both the science and technology education classrooms. Those students who are very diligent workers and well-behaved may receive a less critical eye during the evaluation of their portfolios. I would like to believe that this is never the case, but I suspect that one is unwittingly biased simply by viewing the name on the portfolio. This situation has led me to not viewing the name of the student when I evaluate assignments, labs, exams, and portfolios. Although this may seem somewhat extreme on the surface, I feel that it is the only way in which the halo effect can be minimized. Yet, I still find myself somewhat more critical on labs or portfolios that include difficult to read penmanship. Is it possible that teachers are inherently at risk for the halo effect in the evaluation of their students? Yes, either consciously or unconsciously, which leads one to appreciate the importance of recognizing the erroring potentials of the halo effect.

Logical error refers to a condition in which the examiner "mistakenly deduces more or less correlation between two distinct dimensions than actually exists." (Priestly, 1982, p. 128) Priestly shares an example of a teacher assuming that high intelligence and high academic achievement go hand-in-hand. Conversely low academic achievement would be perceived as correlating with low intelligence. Logical error can influence a judge's evaluation of students which he/she views at differing levels of intelligence and knowingly or unknowingly

scores relative to perceived inteiligence as opposed to actual performance.

The key to improving reliability in the performance assessment process is in the training of examiners and the utility of structured scoring rubrics:

> "Careful training of observers, careful planning and development of the observational assessment, and an analysis of inter-rater reliability can solve most scoring problems. Training involves the selection of persons experienced in the particular field, based on selection criteria such as . . . . geographic region, and education level; and it involves actually training the observers, through demonstration and practice, to make accurate observations and record them properly. Careful planning and development of the assessment includes the construction of the observation instrument, to ensure that it is clear, comprehensive, and meaningful; and it includes uniform guidelines for conducting the observation itself. Analyses of inter-rater reliability may be as simple as having two or more judges observe the same performance, then comparing the results; or it may involve statistical correlation's of all ratings across all raters and examinees."
> (Priestly, 1982, p. 128)

Priestly asserts that the training of examiners is a critical ingredient in the quality and consistency of the performance assessment process. Osterlind (1991) shares seven points that should be included in the training of examiners in the performance assessment process:

> "*Scorers should be informed of the context for the assessments (e.g., whether the assessments are part of a state-mandated assessment program or are a portion of a district's evaluation efforts).
> *Scorers should be aware of answers to the "Why" question. . . . "Why are we using this particular educational

> performance assessment?"
> *Scorers should become thoroughly familiar with the assessment.
> *Scorers should learn the scoring criteria, or scoring rubrics, that are to be used in scoring the assessments.
> *Scorers should become familiar with samples of examinee's papers that are exemplary of particular points in the scoring rubric ("anchor papers").
> *Scorers should gain a working familiarity with the logistics of getting the scoring done, for example, learning the scheme for paper flow and the use of the form on which they will record their judgments.
> *Scorers should be aware of what will happen to the scoring forms after they record their judgments."
> (Osterlind, 1991, p. 65-66)

In this particular research project, Osterlind's perspectives on the training of scorers in the performance assessment process formed the collective start for the eight science teachers involved in the development and implementation of performance assessment tasks in an urban composite high school. The eight teachers were in strong agreement with the perspectives of Osterlind by this studies conclusion.

Gipps (1994, p. 103) addressed the challenges facing the reliability of performance assessment tasks:

> "If traditional test development has over-emphasized reliability at the expense of validity, performance assessment in the same way over-emphasized validity at the expense of reliability. This is because the use of performance assessment is part of a move away from highly standardized procedures. However, if performance assessment is to be used beyond the classroom setting for accountability or certification purposes then we must address questions of reliability."

Gipp (1994) continued to summarize the results of varied studies on

inter-rater reliability in performance assessment tasks:

> "A detailed account of the evidence on inter-rater reliability can be found . . . . These studies indicate that inter-judge agreement can be high on performance assessment tasks but that this has to be achieved through careful training of raters and the provision of scoring rubrics."
> (Gipps, 1994, p. 104)

The structure of the scoring rubrics and evaluator training can play a major role in the reliability of the practical assessment procedure. Studies at Western Michigan University (Kruglak and Wall, 1959) related that inter-rater reliability in performance (practical) assessments were quite low yet could be improved with the implementation of more clearly delineated scoring rubrics:

> "It has been found that the Hoyt reliability coefficients of performance tests were low to moderate but that their values could be increased substantially by subdividing the scoring of each item into several parts. The low reliabilities of performance tests are probably due to their relatively small number of items that can be administered in one session and the great variety of skills and abilities involved."
> (Kruglak and Wall, 1959, p. 26-27)

A study by Le Mahieu (1993) investigated the Pittsburgh portfolio program in regard to inter-rater reliability:

> The Pittsburgh portfolio program used only external assessors who went through a rigorous training and there was a regular checking procedure which led to re-training if acceptable judgments fell below a certain level. Inter-judge reliabilities (IRR) here were at a consistently high level of 0.90. The Vermont portfolio program by contrast used teachers in the schools to carry out the rating with less rigorous training and no checking procedure and attained IRR levels of between 0.34 to 0.43 for the unstandardized

> writing assignments . . . . What emerges clearly from these two developments is that with standardized performance assessment, clear rubrics and training for markers, and exemplars of performance at each point or grade, levels of IRR (inter-judge reliabilities) can be high.
> (Gipps, 1994, p. 104)

This research suggested that the training of the raters and the clear delineation of the scoring rubrics were essential ingredients in the recipe for reliable performance assessment. Gipps (1994, p. 104) shares the conclusions of the research in performance assessment conducted by Linn (1993):

> "With careful design of scoring rubrics and training of raters, the magnitude of the variance components due to raters or interactions of raters with examinees can be kept at levels substantially smaller than other sources of error variance."

The training of the raters/judges/moderators is an important variable that comes into play in the reliability of performance assessment tasks. I am currently training as a flight instructor, and a portion of this training involves assessing the inflight practical abilities of the prospective private or commercial pilot. The scoring rubrics supplied by the Ministry of Transportation (Aviation) Canada, complete with descriptive exemplars, facilitated increased inter-rater reliability of the flight examiners. Moreover, a substantial portion of the flight instructor's training is geared toward increasing the consistency of teaching and inflight evaluation techniques.

Generalizability of Performance Assessment

Generalizability is a term used interchangeably with external validity. External validity pertains to the transferability or generalizability of the results of the experiment, test and so on. Referring to performance assessment, one must ask whether the results of one's performance examination imply with some degree of correlation the abilities of that particular examinee in another task or set of tasks in the tested domain (Gipps, 1994, p. 105). Do the abilities and skills demonstrated in a certain performance assessment accurately reflect that persons abilities and skills overall in that domain of study? Gipps (1994) views performance assessment as problematic in the area of generalizability:

> "Generalizability is a particular problem for performance assessment, since direct assessments of complex performance do not generalize well from one task to another (because performance is heavily task dependent) we cannot take performance on one task to imply that the student could do other tasks in the domain. This is in any event a problem for performance assessment but is more serious if the performance assessment is to be used for anything other than formative, classroom-based purposes. This task specificity is compounded by limited sampling from a domain and the difficulty then of generalizing from the performance to the whole domain."
> (Gipps, 1994, p. 105)

Strategies to deal with the perceived generalizability problems in performance assessment include the use of a larger number of performance skills and problem solving strategies in the design of the

particular performance assessment tasks (Linn, 1993). This problem (generalizability) could be compounded by a teacher's "teaching to task" based on prior knowledge of the material covered on a province/statewide performance exam. In such a situation, to what degree of confidence could the assessment truly relate the learner's appreciation of that (tested) domain in science? In defense of performance assessment, the realities of "teaching to test" in the traditional paradigm of multiple-choice exams are a real issue as well.

Gipps (1994) shares the views of Haertel (1993) on the issue of generalizability in performance assessment and the potential for its analysis on four levels:

> "First, replicable scoring of a single performance (Can we score a single instance of a task in a consistent way?)
> Second, replicability of a specific task (Does the same performance task have a constant meaning across times and places?)
> Third, generalizability across tasks which are presumed to be assessing the same construct (Can we generalize across parallel tasks?)
> Fourth, generalizability across heterogeneous task domains (Can we generalize across tasks that are not parallel?)"
> (Gipps, 1994, p. 107)

The first level depends on the design of the scoring rubrics and the quality and training of the evaluators (Gipps, 1994). As discussed earlier, replicability of scoring in performance assessment has been quite high within exemplars of trained raters and clearly delineated scoring rubrics. The second level, replicability of a specific task, is affected by

the "administration of the task, what the teacher is allowed to say, time allowed, what constitutes coaching, ancillary abilities, reading ability or listening comprehension" (Gipps, 1994, p. 107). Haertel (1993) asserts that generalizability across parallel tasks depends on "the involvement of ancillary abilities and on antecedent instruction; even in a tightly constrained situation in which parallel tasks are kept similar, it is difficult to make two tasks function the same way." (Gipps, 1994, p. 108) In terms of generalizability across tasks that are not parallel, research indicates that performance assessment is "domain specific and that variations in context greatly influence performance" (Haertel (1993) in Gipps, 1994, p. 108). The generalizability issue thus poses a challenge for the designers of performance assessment and illustrates the importance of limiting the scope of extrapolation of the specific performance assessment items "across heterogeneous domains" (Gipps, 1994, p. 108).

Content Validity

Content validity in the practice of performance assessment is seen as a critical factor in the success of this mode of assessment:

> "It is generally agreed that content validity is of paramount concern in criterion-referenced (performance assessment) measurement. . . . A criterion-referenced test may be presumed content valid if all test items are carefully derived from the required performances, conditions, and standards specified in the objectives and if the sample of test items appropriately represents the objectives. . . . The process of

> determining performance criteria on the basis of information obtained directly from job required skills establishes content validity, that is, performance tests that are derived from appropriate task analyses often provide the best available measure of behavioral objectives."
> (Swezey, 1981, p. 149)

In establishing content validity in a specific performance assessment task, one must review the objectives and job-required skills in order to ensure that test items do indeed cover the objectives of a particular learning module. A science teacher must ascertain whether or not the specific aspects being evaluated in his/her performance assessment task indeed reflect the objectives and skills of the curriculum for that particular unit of study. Moreover the skills demanded by that particular test must reflect actual practice. Hence the word "authentic" that is often associated with performance assessment. "From an absolute perspective, a content valid test will demonstrate whether an examinee performs an objective to the required standards or not." (Swezey, 1981, p. 150-151)

Swezey (1981) shares two steps which assist in checking for content validity in the administration of a particular performance assessment task:

> "First it must be determined that objectives have been properly derived from adequate task analyses that prescribe clearly what an examinee must do or must know in order to perform the task under examination.
> Second, each item must be carefully evaluated against its associated objective to ensure that the performances,

> conditions, and standards specified in the item are the same as those required by the objective.
> *If both checks are affirmative, a test (performance assessment task) is content valid."
> (Swezey, 1981, p. 151)

One can appreciate the value of a cooperative science department which could collectively identify and establish performance assessment tasks based on the specific learning objectives mandated for a particular course. Through the employment of a number of qualified professionals, one garners a greater audience with which to check that a particular performance assessment meets the required performances, conditions, and standards specified in the course objectives. (Swezey, 1981, p. 149)

Performance Assessment in Medicine - Lessons Learned

In Educational Researcher, Volume 24, Number 5, June/July 1995, I came across the article "Performance-Based Assessment: Lessons From the Health Professions" by David Swanson, Geoffrey Norman, and Robert Linn. This article summarizes the lessons learned by the health profession in written clinical simulations (patient management problems), computer-based clinical simulations, oral examinations, and standardized patients (live simulations). These lessons included:

> "*Lesson 1: The fact that examinees are tested in realistic performance situations does not make test design and domain sampling simple and straightforward. Sampling must consider both context (situation/task) and construct (knowledge/skill) dimensions, and complex interactions are

present between these dimensions.

*Lesson 2: No matter how realistic a performance-based assessment is, it is still a simulation, and examinees do not behave in the same way they would in real life.

*Lesson 3: While high-fidelity performance-based assessment methods often yield rich and interesting examinee behavior, scoring that rich and interesting behavior can be problematic. It is difficult to develop scoring keys that appropriately reward alternate answers that are equivalent in quality, both because of poor consensus on scoring keys and because of scoring artifacts resulting from variation in response style.

*Lesson 4: Regardless of the assessment method used, performance in one context (typically, a patient case) does not predict performance in other contexts very well. In-depth assessment in a few areas results in scores that are not sufficiently reproducible for use in high-stakes testing.

*Lesson 5: Correlational studies of the relationship between performance-based test scores and other assessment methods targeting different skills typically produce variable and uninterpretable results. Validation work should emphasize study of threats to the validity of score interpretation, not general relationships with other measures.

*Lesson 6: Because performance-based assessment methods are often complex to administer, multiple test forms and test administrations are required to test large numbers of examinees. Because these tests typically consist of a relatively small number of independent tasks, this poses formidable equating and security problems.

*Lesson 7: All high-stakes assessments, regardless of the method used, have an impact on teaching and learning. The nature of this impact is not necessarily predictable, and careful studies of (intended and unintended) benefits and side-effects are obviously desirable but rarely done.

*Lesson 8: Neither traditional testing nor performance-based

> assessment methods are a panacea. Selection of assessment methods should depend on the skills to be assessed, and, generally use of a blend of methods is desirable." (Swanson, Norman, & Linn, 1995, p. 6-11)

The lessons learned by the medical profession in its many decades of actual implementation of performance assessments are undoubtedly of value to secondary science and technology educators who are attempting this mode of assessment.

A Need For Performance Assessment in Science?

The National Assessments of Education Progress (NAEP) in the United States has the central duty to collect data to summarize the state of education nation-wide. National testing at the grades 3, 7 and 11 levels in 1986 assessed four skill areas: "sorting and classifying, observing and formulating hypotheses, interpreting data, and designing and conducting an experiment." (Doran, Lawrenz, and Helgeson, 1994, p. 406) As a result of the NAEP study in 1986 conducted by Educational Testing Service, the following results were shared:

> "Evidence from NAEP and other sources indicates that both the content and structure of our school science curricula are generally incongruent with the ideals of the scientific enterprise. By neglecting the kinds of instructional activities that make purposeful connections between the study and practice of science, we fail to help students understand the true spirit of science . . . . In limiting opportunities for true science learning, our nation is producing a generation of students who lack the intellectual skills necessary to assess the validity of evidence of the logic of arguments, and who are misinformed about the nature of scientific endeavors. The NAEP data support of growing body of literature urging

> fundamental reforms in science education - *reforms in which students learn to use the tools of science to better understand the world that surrounds them."*
> (Mullis & Junkens, 1988, p 17 in Doran, Lawrenz, and Helgeson, 1994, p. 406)

The NAEP further concluded that educators must put in the extra effort and time to include a greater amount of hands-on learning and testing experiences in the classroom. The NAEP stated that:

> "Although managing equipment and training administrators requires ingenuity and painstaking effort, *conducting hands-on assessment is feasible and extremely worthwhile.* The school administrators, teachers, students, and consultants were all very enthusiastic. The students found the materials engaging, and the school staff and consultants were more than supportive in encouraging further use of these kinds of tasks in both instruction and assessment."
> (NAEP, 1987b, p. 7 in Doran, Lawrenz, and Helgeson, 1994, p. 407)

Although the time requirements for the successful design and implementation of performance assessment tasks in secondary science programs are indeed significant, we as educators must go the extra distance to ensure the bridge between theory and practice in science is alive and well in our students' science consciousness.

Chapter 4

Research Strategy and Methodology

Introduction:

The intents of my research project were focused in the examination of practicing teachers' perceptions, abilities, and reactions of performance assessment tasks in the secondary education science environment. This research project involved analysis of the inter-rater reliability demonstrated by these educators while attempting a variety of performance assessment tasks with sample groups of students. Face validity of each set of performance assessment tasks was also investigated by this particular group of secondary education science teachers.

The issue of change, with reference to the educators involved in this project, was addressed. Initial impressions of these particular educators on the topic of performance assessment was shared and compared with their impressions after completion of a battery of performance assessment tasks with sample groups of students. This research provided a rich and detailed "real-world" look at teachers attempting performance assessment activities in the context of an urban composite high school.

The spark that ignited my research vigor was founded in the experience of designing and teaching a set of innovative secondary

applied physics courses at high school level in rural Alberta. These applied physics courses were given the title "High Tech" and offered at the grade 10, 11 and 12 levels (Wozny, 1988). The domains of study in these courses included lasers, robotics, electricity, electronics, pneumatics, hydraulics, alternate power technology, computer assisted design, and aviation/aerospace. I wrote the curriculum for these three courses initially in 1985 and began teaching them in 1989 as a pilot project for Alberta Education. The primary focus of the High Tech courses was to provide students with a hands-on experimental approach to learning which would reinforce the academic skills of mathematics and physics. The instructional model that I designed for these courses incorporated:

1) reinforcement of academic theory,
2) application of theory to real life technological settings,
3) hands-on experimentation with state of the art technological samples,
4) creative problem solving opportunities. (Wozny, 1988)

While teaching these courses, I found myself using an assessment system which included performance assessments (portfolios, creative hands-on problem solving, practical skill exams) as well as the traditional pencil and paper exams. I experienced this assessment system in my training as a commercial pilot, which included both theoretical paper-based exams and hands-on practical flying exams. My

flight instructor's training further developed practical examination delivery. Performance assessment refers to the process of *assessing actual psychomotor (hands-on) manipulation of certain variables and/or problem solving within the view of a moderator.*

My experiences in the utility of a composite assessment system in the High Tech program and my commercial pilot training (composed of traditional pencil and paper exams coupled with performance inflight evaluations) provided the motivation to pursue my doctoral research into the process and practice of performance assessment in physics. In addition to my research into performance assessment, I composed a short and concise teacher booklet and multi-media presentation on integrating performance assessment strategies in the classroom based on a selected review of related literature, as well as my classroom experiences and the observations of those participants in this study. This booklet and multi-media presentation are in the appendix of this dissertation.

Statement of the Problem:

The intents of my research are grounded in the examination of teachers' initial and post-study perceptions of the feasibility of the performance assessment process in secondary science coupled with an analysis of inter-rater reliability and face validity evidenced in the actual trials involving these particular instructors. The following questions

summarize the problems addressed by this research project:

1) What are science teachers' initial perceptions of the feasibility in the performance assessment process in the secondary education science classroom?

2) What are the inter-rater reliabilities amongst a group of secondary education science teachers involved in actively creating and completing a battery of performance assessment tasks with small groups of students?

3) What are the participating teachers' thoughts on the face validity of the particular performance assessment tasks enacted in this study?

4) What are the participating teachers' thoughts on the feasibility of performance assessment tasks in the classroom after several trials involving sample groups of students?

This research project also included a brief teacher booklet and multimedia presentation on the integration of performance asssessment in the science and technology education classroom. It was my intention to provide the practicing teacher with a useful tool in appreciating the dynamics and potentials of the performance assessment process in education. One of the problems that often besets researchers is making their research available to the practicing teacher. By composition of a handbook coupled with a user-friendly (Microsoft Power Point) presentation on the fruits of this research, I hope to allow for relatively convenient access by interested audiences.

Timetable:

1) Proposal submitted and approved (including ethics review) by supervising committee - August, 1995.

2) Permission obtained from selected composite high school for research interaction with science instructors - May, 1995.

3) Establishment of connections with 8 science instructors to be utilized as case study/action research participants in this dissertation project - May, 1995.

4) Research project initiation and data collection - September, 1995 to June, 1996.

5) Dissertation write-up and possible defense - July, 1996 to April, 1998.

Scope and Delimitations of this Study:

This research project involved eight secondary education science teachers currently employed at an urban composite high school. The scope of this study included the investigation of these science teachers' initial perceptions on the feasibility of the performance assessment process coupled with the actual attempting of a battery of performance assessment tasks created by these individuals. An analysis of the inter-rater reliability and face validity evidenced in this study was included. This study also integrated an analysis by the participants (eight science teachers) of their reflections on the feasibility of performance assessment

tasks in the science classroom after completion of the sample performance tasks with the particular groups of students.

This research project also included a brief and concise booklet "A Practicing Teacher's Ten-Minute Guide Into Performance Assessment in Science and Technology Education." I also included a brief Power Point multimedia presentation providing the practicing teacher with concepts which may increase his/her knowledge and understanding of the performance assessment process and potentials for integration of such assessment paradigms in his/her particular educational situation.

Limitations:

The eight science instructors who volunteered for this research project are not considered to be a representation sample of secondary physics teachers, but simply a group of individuals willing to volunteer their time and services to this particular study. I was fortunate to gain the input of an entire science department of a medium sized (approximately 700 students) composite high school. The students who participated in the performance assessment tasks were after-school volunteers who consisted of wide range of academic abilities from strong to weak.

This study described the perceptions of these eight science teachers on the feasibility of the performance assessment process both before and after the development and administration of the performance

assessment tasks. This group of teachers designed the performance assessment tasks, implemented them with small groups of volunteer students and shared results for inter-rater reliability analysis. Face validity judgements by the participating teachers were also included for each of the performance assessment tasks employed.

The teacher handbook on performance assessment is a brief and concise "user-friendly" document that shares some theories and examples of performance assessment strategies. I purposefully wrote the booklet using plain language so as to minimize the intimidation factor, especially for new teachers. This document will hopefully assist practicing teachers with comprehension of this evaluation paradigm. A floppy disk is available in the pouch of this dissertation which contains the accompanying multimedia presentation written with Microsoft Power Point.

Definition of Terms:

action research: a collaborative research system involving the on-going processes of planning, action, observation and reflection.

> "Action research is . . . . a form of collective self-reflective inquiry undertaken by participants in social situations in order to improve the rationality and justice of their own social or educational practices, as well as their understanding of these practices and the situations in which the practices are carried out"
> (Kemmis and McTaggart, 1988)

case study (qualitative): This research methodology involves the holistic

appreciation of the variables involved in a particular situation.

> "In a case study the investigator attempts to examine an individual or unit in depth. The investigator tries to discover all the variables that are important in the history or development of the subject. The emphasis is on understanding why the individual does what he or she does and how behavior changes as the individual responds to the environment."
> (Ary, Jacobs, and Razavieh, 1990)

constructivism: Appreciation of the variables and dynamics which involve a particular issue within the context of its environment: "a way of interpreting and making sense of a variety of phenomena. Constructivism constitutes a framework within which to address situations of complexity, uniqueness, and uncertainty" (Cobb, Wood, and Yackel, 1991). "The essential way of knowing the real world is not directly through our senses, but first and foremost through our material or mental actions" (Sinclair, 1990).

performance assessment: Performance assessment (often referred to as practical assessment) is a process of evaluation involving the learner in the solution of problems and/or practical skill challenges within the access of a judge/moderator.

> "Performance assessment is defined as a systematic attempt to measure a learner's ability to use previously acquired knowledge in solving novel problems or completing specific tasks. In performance assessment, real life or simulated assessment exercises are used to elicit original responses which are directly observed and rated by a qualified judge."
> (Stiggins and Bridgeford, 1982)

portfolios: Samples of a student's or group of students' work in a collecting medium such as a scrapbook, log book, computer file, etc.

Significance of this Study:

This research is extremely important in providing practicing educators the opportunity to appreciate the varied and complex factors that interplay in the integration of performance assessment tasks in the science classroom. This need is further extenuated by the recent enactment of curriculum changes by provinces and states, that include aspects of performance assessment in the evaluation structure of science and technology education courses at the high school level. Alberta Education is currently developing a set of performance assessment tasks for utility in junior high and senior high science programs. In the United States the National Assessment of Educational Progress (NAEP) has been involved internationally in reviewing the potentials of integrating performance assessments in the assessment structure of the American classroom. In Connecticut significant efforts have been placed into the development and delivery of performance assessments in the classroom:

> "Connecticut incorporated performance assessments into its statewide testing programs in art and music, business and office education, English language arts, foreign languages, industrial arts, mathematics, and science. In the late 1980's and early 1990's, other states, California and Vermont, for example have been piloting a variety of approaches that incorporate performance assessments into science, mathematics, language arts, and social studies. Judging from the size of audiences at recent national meetings on

> performance assessment, interest is growing."
> (Boykoff Baron, 1991)

The process of integrating performance assessment in the classroom is very active in Great Britain, Australia and the Netherlands:

> "Examples include the new national curriculum in Great Britain, innovative mathematics curriculum and teaching programs in Australia, and the national mathematics project in the Netherlands. In this view, assessment is integrated naturally into the curriculum, and the assessment itself models good instruction. Thus a performance assessment becomes the culminating activity of a unit and provides opportunities for students to synthesize their knowledge, make connections, and deepen their understanding of major concepts. The assessment creates situations that are intended to foster the development of deeper levels of understanding. This new view of performance assessment blurs the edges among assessment, curriculum and instruction." (Boykoff Baron, 1991)

The dynamics of implementing performance assessment tasks in the traditional pencil and paper assessment paradigm of many a classroom will undoubtedly encounter a myriad of challenges which will affect its potentials for success. The process of resocializing students, teachers, and administrators to the complexities of performance assessment will require the input of researchers disseminating teachers perceptions and experiences about the performance assessment process. The booklet and multimedia presentation describing the fruits of this particular research project will hopefully facilitate a user-friendly means of sharing information with fellow educators.

Methodology and Procedure:

This research project was based in the study of science teachers' initial perceptions, abilities, and reactions about performance assessment tasks in the secondary education science environment. Inter-rater reliability and face validity of each particular performance assessment task designed and implemented was investigated and analyzed by the research group science teachers. This examination coupled with a selected literature review of the performance assessment process in secondary education science formed the basis for the composition of this research project.

As mentioned previously, I used aspects of case study and action research qualitative research perspectives grounded on the argument that:

> "human behavior is always bound to the context in which it occurs, that social reality (for example, cultures, cultural objects, institutions, and the like) cannot be reduced to variables in the same manner as physical reality, and that what is most important in the social disciplines is understanding and portraying the meaning that is constructed by the participants involved in particular social settings or event. Qualitative inquiry seeks to understand human and social behavior from the insider's perspective, that is, as it is lived by participants in a particular social setting (for example, a culture, school, community, group, or institution)."
> (Ary, Jacobs, and Razavieh, 1990, p. 445)

The context of the composite high school science classroom, coupled with the actual teachers responsible for the instruction and

evaluation in that particular facility, provided a natural setting in which I related a holistic picture of teachers' perceptions and practice on the feasibility, reliability, and validity of the performance assessment process. I felt it was imperative to work with practicing teachers in order to gain a realistic, non-contrived view into the real world of science performance assessment tasks in practice. This is consistent with the qualitative research paradigm due to the reality that "what can be learned in a particular setting depends on the nature and types of interactions between the inquirer and the people and setting and those interactions are not fully predictable, and because important features in need of investigation cannot always be known until they are actually witnessed by the investigator." (Ary, Jacobs, and Razavieh, 1990, p. 448)

Aspects of the case study research methods used in this project allowed for an in-depth analysis and description of this particular research perspective into performance assessment. Sharan Merriam in her book, Case Study Research in Education, (1988) relates a view of case study research in which:

> "A case study is an examination of a specific phenomenon such as a program, an event, a person, a process, an institution, or a social group. The bounded system, or case, might be selected because it is an instance of some concern, issue, or hypothesis . . . . one would study it to achieve as full an understanding of the phenomenon as possible . . . . case study seeks holistic description and explanation." (Merriam, 1988, p. 9 - 10)

Further to this view of case study research Merriam (1988, p. 11 - 12) shares four characteristics which are integral aspects of the qualitative case study which I employed in the research methods of this particular project:

1) *Particularistic* - case studies are directed towards a particular event, happening, or practice. Case studies "concentrate attention on the way particular groups of people confront specific problems, taking a holistic view of the situation. They are problem centered, small scale, entrepreneurial endeavors" (Shaw, 1978, p. 2).

In this study I worked with the entire science department of a particular composite high school. In this high school we collectively investigated the potentials of performance assessment tasks in the science classroom. The strength of this approach is based in the actual working environment of the practicing science teacher.

2) *Descriptive* - a very detailed and contextually accurate description of the aspect being studied is provided along with the surrounding environment. The interpretive or hermeneutic appreciations of the researcher and participants are shared so as to allow the reader to appreciate the complexity and dynamics of the particular situation.

By allowing the teacher-participants of this study to share in the planning and procedure of this research project, we are allowed a view into the practicing teachers' assumptions and practice in the domain of

performance assessment tasks. We are also privy to the inner workings of a composite high school over a one-year duration.

3) *Heuristic* - the case study methodology seeks to inform the audience of the deep understanding of the research topic. "Previously unknown relationships and variables can be expected to emerge from case studies leading to a rethinking of the phenomenon being studied. Insights into how things get to be the way they are can be expected to result from case studies." (Stake, 1981, p. 47)

In this particular research project, we are able to peer into the perspectives and practice of an entire science department of a medium sized composite high school over an entire school year. Moreover, we gain an appreciation of the assumptions of these teachers in regard to performance assessment, both in theory and practice. Future educators, who have the opportunity to review this study, will gain insight into the practicing teachers' perspectives and pedagogy in the domain of performance assessment tasks. The detailed interviews with the participants in this study provided a realistic and detailed perspective from the real world teacher into the perceived strengths and weaknesses of the performance assessment process.

4) *Inductive* - the utility of inductive reasoning dominates in the qualitative case study methodology. The process of discovery through research is often the focus in the case study in contrast to the

experimental research approach where verification/falsification of the hypothesis is the central aim.

The qualitative case study approach, coupled with aspects of action research, provided a means of investigating and understanding of performance assessment in science which co-emerged through the research process. This allowed for a rich and holistic appreciation of the performance assessment activities and process in the real world of the practicing secondary science teacher.

I was drawn to this particular mode of research (case study and action research) due to the potential for holistic appreciation of the subject, performance assessment, and the surrounding environment. The audience was given the opportunity to appreciate the *meaning* of this research experience in the context of the urban composite high school. This research allowed for the growth and development of the researcher and subjects within the natural setting of the science classroom.

Chapter 5 - Teachers' Initial Perceptions on the Feasibility of Performance Assessment

Each of the eight science teachers involved in this particular research project shared their thoughts in response to a written questionnaire asking for their input on the feasibility of integrating performance assessment tasks in their science program. This question was asked to ascertain the initial impressions of the teachers on the performance assessment process. One must appreciate that this questionnaire was completed at the beginning of this research project without previous sharing amongst these science teachers on the topic of performance assessment.

Teacher A (Ms. O.): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- Performance assessment tasks are quite time consuming; time factor would be a big deterrent for me.
- They (performance assessment tasks) would be feasible with small classes, less than ten or twelve students.
- If all students in a given class need to be assessed on the same tasks the assessment would prejudice in favor of the students in the second or latter half of the class (if the task was repeated by each successive student while others were in the room).
- The feasibility would increase in a lab setting where I could walk

around and have students demonstrate certain skills while others were working on the lab; for example titration techniques.

- Using performance assessment as a part of an exam would see the need for each student to be scheduled. This would be very time consuming and students could communicate with their friends as to what the tasks specifics are. (Is the assessment then valid?)

Teacher B (Mr. A): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- Could be done but might involve more time than is practical. The program might have to be analyzed to determine the tasks that criteria could be attached to mark easily.

- One of the problems is that the activities are good but the concepts in many cases are quite subtle.

- Our courses are already so activity oriented. I am not sure that science should be perceived as a series of experiments. Rather experiments should be born from theories, observation, or problems.

- Some students may benefit from this approach but I am not sure that all students should be subjected to this. I suppose that I am of the old school. Does performance necessarily mean that a student understands. If a student understands will he or she necessarily perform well physically?

Teacher C (Ms. K.): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- Very difficult, especially with large classes, to complete frequently unless another supervisor is available to guide students when not participating in test.

- Feasible if tasks assessed are not too lengthy or complex.

Teacher D (Ms. A): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- We should be moving towards performance-based assessment since indications from Department of Education show this will be "expected" as evaluation in the future.

- Teachers need time to develop performance-based assessment tasks; to try them and refine in class.

- Collaboration amongst science teachers would help generate ideas and edit, or draw up alternatives.

Teacher E (Mr. H): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- I believe it is difficult with a larger group to be fair and consistent in marking unless you can watch all students perform the same tasks. With a smaller class it may be possible.

- Checklists of a particular task as useful in assessing certain tasks such as titration, filtration, etc.

- Having said this, even some evaluation of lab techniques, fair or not, makes students more aware of their lab techniques. If they are aware you are watching they make more of an effort because it is "for marks."
- I have made minimal use of this assessment technique in the past and only at junior high level. It becomes more difficult with large classes at the 20 to 30 levels.

Teacher F (Ms. J): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- I think it is very practical though time consuming.
- This idea of it (performance assessment) would very much profit our task of effective evaluations.
- My problem with the idea of performance assessment is time. Time wise it would be difficult to schedule, but I think it could be set up.

Teacher G (Mr. D): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- I think if you have a class with a small number of students, less than 20, and they are well behaved, performance assessment is feasible and really good especially for those students who don't fair so well on written tests.
- I don't believe it is feasible, without extra help from another staff member, to do performance assessment in a normal grade nine science class. As you are checking out the performance of 1 to 3 students,

someone has to be looking after the rest of the class.

- Also, if the assessment is spread over days, weeks, or months, one student will tell the next, making it unfair for the earlier ones. This would require changing the task everyday, hopefully maintaining the same degree of difficulty.

I think performance assessment is great in theory but I'm not sure about how great it is in practice.

Teacher H (Mr. P): *What are your thoughts on the feasibility of integrating performance assessment tasks in your science program?*

- I am initially concerned with the time necessary to create, refine and then deliver this mode of assessment, especially when dealing with large classes and a busy schedule with minimal preparation time.

- The consistency (reliability) of my assessments of student performance would be of concern, appreciating the change in perspective as one assesses a particular task over a long period of time. How can one be certain that you are indeed consistent from student to student?

- How/What would be the optimal scoring rubrics to allow for clear and consistent marking by the prospective evaluator?

- How can one be certain that this particular performance assessment task is indeed valid within the context of its delivery?

- What training process could be implemented in order to facilitate accessible instruction to prospective teachers interested on the

integration of performance assessment strategies in the science classroom?

Summary of Teachers' Thoughts on the Feasibility of Integrating Performance Assessment Tasks in the Science Classroom

The issue of time comes up in every teacher's initial thoughts on the performance assessment process. It takes time to design, set-up and complete performance assessment tasks which may not be viable in a busy schedule coupled with large classes. One must appreciate that this question was posed to these teachers at the very start of this research project, before any of us had collectively met and discussed the issue of performance assessment in the science classroom. Many teachers in this research project commented on the pressure to achieve high results on the diploma exams (traditional pencil and paper) at the grade 12 level, which often left little time for performance assessments. In order to cover the entire curriculum for a grade 12 science course, one tended to focus on diploma exam test items.

The second major issue shared by these teachers was that of class size. If one were teaching very large classes (greater than 30 students) would it be realistic to implement performance tasks on an individual basis? Most teachers believed that performance assessment tasks were feasible with smaller classes (up to 20 students), but difficult with larger numbers.

The issue of fidelity in the performance assessment tasks was shared by a majority of the teachers. If students were to be made aware of the content of a particular performance assessment task well before the actual exam date, it could enable them to practice that skill, perhaps giving a false impression of the students' problem-solving abilities. This would be particularly crucial in performance assessment tasks that involved the solution of a novel problem, which required logic and innovation.

Reliability of the performance assessment process was shared as a major concern, especially in a large class where a teacher may unintentionally change the "ruler" of assessment as the day wears on. How could one ensure that the assessment of hands-on skills would be consistent from student to student?

The development of performance assessment tasks was generally seen as an involved process requiring the collective inputs of the science department as a team. One teacher remarked on how science teachers in Japan's public school system get their preparation periods at the same time as their peers which allows for collective development of pedagogical materials such as performance assessment tasks.

The concern for having to integrate performance assessment activities into the science classroom was shared by some of the interviewed teachers. One teacher in particular noted that performance

assessments were now expected by us in certain science activities and she was quite concerned about her ability to do such assessments in lieu of her busy schedule. Moreover, some teachers felt uneasy about the performance assessment process due to lack of training and experience in the design and implementation of this assessment paradigm.

Chapter 6 - Analysis of Performance Assessment Tasks Completed by Teacher Research Group

A total of five performance assessment activity sets were designed and utilized by the group of eight science teachers involved in this particular research project. These tasks were implemented during the 1995/96 school year at an urban composite high school in central Alberta. The performance assessment tasks are written in the chronological order in which they were given.

Performance Assessment 1 - Basic Electronics

Our research group collectively decided on an initial set of performance assessments which involved four specific tasks to be completed by grade nine students individually. These tasks were directly related to the grade nine science curriculum (Alberta Program of Studies, 1993). It should be noted that the eight teachers involved in this study were very cooperative in this process of designing the tasks. As the author of this study, I was seen as the leader of the research team, yet the participating teachers were very willing to provide input and constructive criticism as we developed our first performance assessment activity.

Task 1 challenged the student to draw a schematic diagram of one lamp connected in series with two lamps in parallel, which were all to be connected to a six volt direct current source. This schematic had to include one switch to turn on/off all three lamps and include two more

switches to turn on/off either of the two lamps in parallel.

Task 2 involved the student constructing this schematic with provided materials. The student was to operate the circuit and explain the differences between the series portion and parallel portion of the circuit.

Task 3 of this performance assessment had the student use the voltmeter of his/her choice to measure the voltage drop across the lamp in the series portion of the circuit.

Task 4 involved the calculation of energy used by an electrical direct current circuit which drew 0.15 amperes of current at 6.0 volts over a 2.0 hour time period.

Three grade nine science students volunteered to attempt these performance assessment tasks individually while being observed and evaluated by the eight science teachers. Scoring rubrics developed by the research group of teachers were distributed to the students before the actual assessment process was initiated. The scoring of the performance assessment tasks in electronics were based on a 0 to 4 point scale:

4 points - Skill demonstrated with full competence with no assistance.

3 points - Skill demonstrated with high level of competence with occasional guidance.

2 points - Skill demonstrated with mid-level competence with

occasional guidance.

1 point - Skill demonstrated with fair to poor competence with frequent guidance.

0 point - Absence of skill demonstrated coupled with need for continual guidance.

Three grade nine science students volunteered for this particular set of performance assessment tasks. Two of the students had science marks in the 60% to 70% range and one student had an average of over 85%. Five out of the eight science teachers at this composite high school were available to assess each of these students for the entire duration of this particular set of performance assessment tasks.

Before the students actually attempted these tasks, the teachers who designed this set of performance assessments in the domain of grade nine basic electricity collectively reviewed what this activity entailed coupled with brief explanations of correct wiring procedures.

It was decided by the collective group of teacher/examiners that we would only have the students perform tasks one and two of this particular set of performance assessments in electricity, due to the time constraints which were placed at approximately 15 minutes per student for the duration of the performance assessment procedure.

Results:

The following scores represent the results attained by each student and the score allocated by each teacher for each specific task.

Table 1A
Student 1 - Tom

Teacher	Task 1	Task 2
	Score /4	Score /4
Mr. P	2	2
Ms. A	2	2
Mr. D	1	1
Ms. K	1	2
Mr. A	2	3

Task 1 average = 2.2
Task 2 average = 2.6
Standard Deviation of Scores on Task One = 0.55
Standard Deviation of Scores on Task Two = 0.71

Table 1B
Student 2 - Jack

Teacher	Task 1	Task 2
	Score /4	Score /4
Mr. P	4	4
Ms. A	4	4
Mr. D	4	4
Ms. K	4	4
Mr. A	4	4

Task 1 average = 4
Task 2 average = 4
Standard Deviation of Scores = 0

Table 1C
Student 3 – Jim

Teacher	Task 1	Task 2
	Score /4	Score /4
Mr. P	3	1
Ms. A	3	2
Mr. D	2	2
Ms. K	3	2
Mr. A	3	2

Task 1 average = 2.8
Task 2 average = 1.8
Standard Deviation of Scores on Task One = 0.47
Standard Deviation of Scores on Task Two = 0.47

Overall average score for student 2 - Jack was 4/4 which correlated to skill demonstrated with full competence with no assistance for task one and two. Interestingly enough, with this particular student, all the teachers were in absolute agreement about the scoring of both activities demonstrated by Jack, the grade nine science student, who had also scored very high (85%) in his written tests in this domain of study. It should be noted that Jack is an above average student with a keen interest in electronics, which includes extensive hobby experience. Jack's father is an electrician and has involved Jack in many electronic projects from a young age.

Summary Of Interrater Reliability For Performance Assessment 1 - Electronic Circuits

The Pearson's Product Moment Correlation Coefficient was used to calculate the inter-rater reliability between each of the examiners. This coefficient is representative of the strength of correlation of one examiner's score with another examiner's score. The range of possible positive correlations are as follows:

1.0 perfect positive correlation
0.8 very strong positive correlation
0.6 moderate to strong positive correlation
0.4 weak correlation
0.2 very weak correlation
0 probably no correlation
(Fitz-Gibbon, Morris, 1978, p. 92)

Table 1R
Inter-rater Reliability Matrix of Five Raters For Performance Activity One
(Correlation Coefficients Pearson's Product Moment)

	Mr. P	Ms. A	Mr. D	Ms. K	Mr. A
Mr. P		0.95	0.81	0.86	0.92
Ms. A			0.94	0.95	0.91
Mr. D				0.96	0.81
Ms. K					0.92
Mr. A					

Overall inter-rater reliability for performance assessment task 1 (Electronic Circuits) using Pearson's Product Moment Correlation Coefficient over 60 paired scores = 0.83

Summary Thoughts On Performance Assessment 1 - Electronic Circuits

This initial foray into the domain of performance assessment was a rousing success. As one can observe from the inter-rater reliability comparisons, this particular performance assessment experience proved to be a reliable mode of assessment in this test case, as evidenced by the overall inter-rater reliability correlation of r = 0.83. I utilized Pearson's Product Moment Correlation in the statistical analysis of inter-rater reliability of this particular performance task which was accurate to a lower limit of 95% confidence interval which was r = 0.75 and an upper limit of 95% confidence interval of r = 0.90. This confidence interval was based on the 60 paired correlation scores resulting from this particular performance assessment task.

Face validity of performance test "Electronic Circuits" was seen as being sound by the group of science teachers, based on its close parallel with curriculum content in the Alberta Science 9 program of studies. Students were expected to have the skills to design and construct both series and parallel circuits at this level of science. Comments from the participating teachers were very positive in regard to the validity of this particular set of performance tasks. The positive results from this first performance assessment activity sparked the implementation of similar performance exams in the participating science teachers' classrooms in which this mode of assessment was not a usual event in the evaluative

schemata of that particular instructor.

Two of the teachers involved in the implementation of this particular performance assessment suggested that we should rehearse the performance assessment in detail before actual utility by our students. It was felt collectively that this rehearsal would ensure that we all view representative performances at each of the various quality levels in order to increase the consistency of our appraisals during actual testing in the classroom.

The science teachers involved in this project, including myself were quite shocked at how reliable this initial attempt turned out to be ($r = 0.83$). Several teachers commented on that they did not think that our initial performance tasks would have such a strong positive correlation.

Performance Assessment 2 - Measuring Density

The second performance assessment session implemented by our research group of science teachers involved the students measuring the density of an irregularly shaped object. This skill is included in the grade nine science curriculum (Alberta Education). The students were provided with a variety of materials and apparatus to complete this challenge, but were not advised which apparatus they were going to use or how to use it. Students were given access to the scoring rubrics before they attempted this particular performance assessment in order to increase their realization that they would be penalized if they seek and

receive guidance from the teachers.

In this particular set of performance assessment tasks the participating science teachers discussed in detail how we would assess the various types of performance illustrated by the examinees. We (science teachers) felt that a clear understanding of the level of competence for the various grade allocations was essential for increased inter-rater reliability. Simply put, we hoped to increase our reliability in this second set of performance assessments by training ourselves as to what delineated each specific level of performance as set by our scoring rubrics.

The scoring rubrics used in this assessment were as follows:

4 Skill demonstrated with full competence, with minimal teacher guidance.

3 Skill demonstrated with high level of competence with occasional teacher guidance.

2 Skill demonstrated with mid-level competence with occasional teacher guidance.

1 Skill demonstrated with fair to poor competence with frequent guidance.

0 Absence of skill demonstrated coupled with need for continuous guidance.

The scoring categories included:

1) Knowledge of density relationship /4

2) Data collection strategies /4

3) Materials and Apparatus skills /4

4) Extra - Error Recognition /1

The apparatus supplied for this particular assessment included a triple beam balance, 2 - 50 milliliter graduated cylinders, water, thread, pliers, beaker, test tube, ruler, pencil, calculator, and paper. Students were advised to describe their knowledge of density before they embarked in the solution of the density of a complex irregular shaped object, which in this case was a threaded bolt.

This performance assessment activity involved seven science teachers and four volunteer grade nine science students. These students were not a representative sample of the students in this particular school, but do provide actual test subjects with which to research the reliability of the teachers' assessments. This performance assessment test was considered to be a valuable tool because these students had just completed study in the determination of density of irregular shaped objects in their science nine courses. The teachers involved in this study were curious to see if these students could share a basic understanding of density and its determination from mass and volume, especially since it should have been quite fresh in their (students') memory.

Results:

Task 1 = Knowledge of Density Relationship /4
Task 2 = Data Collection Strategy /4
Task 3 = Materials and Apparatus Skills /4
Task 4 = Extra (Error Analysis) /1

Table 2A
Student 1 Jim - Scores for Performance Assessment 2 - Density

Teacher	Task 1	Task 2	Task 3	Task 4	Total
Ms. O	1	2	2	0	5
Ms. J	0	4	3	0	7
Ms. A	0	4	3	0	7
Mr. H	1	4	4	0	9
Mr. D	N/A	N/A	N/A	N/A	
Ms. K	0	4	3	0	7
Mr. A	1	2	2	0	5

Average Total = 6.6

Standard Deviation of Scores on Task 1 = 0.55
Standard Deviation of Scores on Task 2 = 1.0
Standard Deviation of Scores on Task 3 = 0.75
Standard Deviation of Scores on Task 4 = 0
Standard Deviation of Scores on Total /13 =1.5

Table 2B
Student 2 Fred - Scores for Performance Assessment 2 - Density

Teacher	Task 1	Task 2	Task 3	Task 4	Total
Ms. O	4	4	3	1	12
Ms. J	4	4	3	1	12
Ms. A	4	4	4	1	13
Mr. H	4	4	4	0	12
Mr. D	4	3	3	0.5	10.5
Ms. K	4	4	3	1	12
Mr. A	2	4	2	0	8

Average Total = 11.4

Standard Deviation of Scores on Task 1 = 0.76
Standard Deviation of Scores on Task 2 = 0.38
Standard Deviation of Scores on Task 3 = 0.69
Standard Deviation of Scores on Task 4 = 0.48
Standard Deviation of Scores on Total /13 = 1.7

Table 2C

Student 3 Brian - Scores for Performance Assessment 2 - Density

Teacher	Task 1	Task 2	Task 3	Task 4	Total
Ms. O	1	1	1	0	3
Ms. J	0	2	2	0	4
Ms. A	0	1	2	0	3
Mr. H	0	2	2	0	4
Mr. D	1	3	2.5	0	6.5
Ms. K	0	2	4	0	6
Mr. A	1	1	2	0	4

Average Total = 4.4

Standard Deviation of Scores on Task 1 = 0.53
Standard Deviation of Scores on Task 2 = 0.76
Standard Deviation of Scores on Task 3 = 0.90
Standard Deviation of Scores on Task 4 = 0
Standard Deviation of Scores on Total / 13 =1.5

Table 2D

Student 4 Bradley - Scores for Performance Assessment 2 - Density

Teacher	Task 1	Task 2	Task 3	Task 4	Total
Ms. O	0	1	1	0	2
Ms. J	0	1	1	0	2
Ms. A	0	2	2	0	4
Mr. H	0	1	2	0	3
Mr. D	1	.5	1	0	2.5
Ms. K	0	2	1	0	3
Mr. A	0	1	1	0	2

Average Total = 2.6

Standard Deviation of Scores on Task 1 = 0.38
Standard Deviation of Scores on Task 2 = 0.57
Standard Deviation of Scores on Task 3 = 0.49
Standard Deviation of Scores on Task 4 = 0
Standard Deviation of Scores on Total / 13 = 0.75

Summary Of Inter-rater Reliability For Performance Assessment 2 - Measuring Density

The Pearson's Product Moment Correlation Coefficient was used to calculate the inter-rater reliability between each of the examiners. This coefficient is representative of the strength of correlation of one examiner's score with another examiner's score. The range of possible positive correlations are as follows:

1.0 perfect positive correlation
0.8 very strong positive correlation
0.6 moderate to strong positive correlation
0.4 weak correlation
0.2 very weak correlation
0 probably no correlation
(Fitz-Gibbon, Morris, 1978, p. 92)

Table 2R
Interrater Reliability Matrix of Seven Raters (Density Tasks) (Correlation Coefficients - Pearson's Product Moment)

	Ms. O	Ms. J	Ms. A	Mr. H	Mr. D	Ms. K	Mr. A
Ms. O		0.98	0.98	0.92	0.91	0.97	0.96
Ms. J			0.96	0.98	0.94	0.98	0.98
Ms. A				0.94	0.84	0.91	0.91
Mr. H					0.91	0.92	0.94
Mr. D						0.98	0.98
Ms. K							0.99
Mr. A							

Overall inter-rater reliability for performance assessment task 2 (Measuring Density) using Pearson's Product Moment Correlation Coefficient over 78 paired scores = 0.87

Summary Thoughts On Performance Assessment 2 - Measuring Density

This performance assessment activity yielded a very strong correlation in both the assessor comparison matrix and overall inter-rater reliability as evidenced by the high correlation, r = 0.87. Application of Pearson's Product Moment Correlation Coefficient with 78 paired scores produced a lower limit of 95% confidence interval which was r = 0.83 and an upper limit of 95% confidence interval of r = 0.92. The large number of 78 paired scores minimizes the error in the calculation of the Pearson's Product Moment Correlation Coefficient as evidenced by the small spread between the upper and lower limit of 95% confidence intervals for this particular inter-rater analysis.

It is quite interesting to note that the inter-rater reliability increased from 0.84 in our first performance assessment activity "Electronic Circuits" to over 0.87 in "Measuring Density," the second activity. It should be noted that, in the second performance assessment activity, the teachers' involved in this research project discussed the scoring rubrics in detail as well as the requirements necessary in the student's performance which would result in a particular grade. It would appear that delineation of the exemplar performances at each assessment level increases the inter-rater reliability of the performance assessment process. As well, we had a greater number of teacher evaluators in this assessment (seven examiners) in comparison to the

first that involved five teachers in "Electronic Circuits". One of the teachers decided to mark in 0.5 increments, although we had initially decided to mark in single digit increments. This teacher felt that the 0.5 increment would allow him greater precision in the assessment process.

Face validity in this second performance task "Measuring Density" was considered very strong by each and all of the seven participating teachers due to its inclusion in the curriculum of the grade 9 science students. This activity reinforces a laboratory activity included in the text for the science 9 program of studies. As well, this skill was indeed tested in the provincial science 9 achievement exam (Alberta) in June, 1996, although on the provincial test it was in multiple choice written form and did not involve the detail and skill analysis which was analyzed in our group assessment process.

Many teachers involved in the second performance assessment project were very positive about the possible inclusion of this mode of assessment in their particular science class, but voiced concerns over the large time requirement to assemble and implement this mode of assessment. Additional concerns were once again voiced about the fidelity of such an assessment process where previously tested students might have the opportunity to share details about the performance assessment to others. One teacher commented that such a situation would not necessarily matter, if the "overall purpose was for this skill to

be in place for all students." In essence this teacher felt quite certain that one could even tell the students quite specifically what their particular performance assessments would involve and then test at a later date. This idea was based in the assumption that students will invest time and effort in those skills and abilities that will have a direct affect in their overall mark. A second teacher voiced the thought that such tests would be of significant value in ensuring that certain skills were mandated in the science students through each particular level. Moreover, this process would hopefully generate more consistency in the skill levels of science students across the province. This is not a trivial concern in light of safety in the setting of the science laboratory.

Performance Assessment 3 - Microscope Skills

The third performance assessment activity initiated and implemented by the research group of science teachers involved the demonstration of proper microscope use and design of a scientific drawing. The object of the scientific diagram was an onion root tip cell, in which the student was challenged to label five cell parts while the microscope was set on medium power.

The scoring rubrics were broken down into two sections:

1) Microscope Use - Five Marks Overall - ½ mark for each of the following behaviors exemplified by the examinee:
- uncover microscope
- turn on light
- put slide on stage

- on low power, use coarse focus
- adjust iris diaphragm
- on medium power, use of fine focus
- return to low power
- removal of slide
- turn off light
- replace dust cover

2) Scientific Diagram - Ten Marks Overall

- title in capitals	- 1 mark
- labels lined up to right of sketch in lower case letters	- 1 mark
- line sketch in pencil (no shading)	- 1 mark
- magnification = objective x ocular = 100x	- 1 mark
- name and date in lower right corner	- 1 mark
- five correctly identified parts - ie. cell wall, cell membrane, cytoplasm, nucleus, nuclear membrane/envelope, nucleoplasm, nucleolus	- 5 marks

The teacher (Biology 20,30) that designed this particular performance assessment task felt that this delineation of the scoring rubrics would aid in the ease and reliability of the evaluation process. Three grade ten students were invited after school into the biology laboratory to test out this particular performance assessment test. Two of the students were very high achievers in science and the other average as measured by their marks over the last term in Science 10.

Although the previous two performance assessment tasks were collectively designed by the research group, it was thought by the research group that one teacher in particular should design the entire performance assessment task. Upon completion of the performance assessment task design, the biology teacher invited the other science teachers to act as examiners after a brief dialogue on the test design and

structure.

The assignment page presented to the students went as follows:

> Demonstrate proper microscope use and make a scientific drawing of an onion root tip cell by labeling five cell parts while on medium power. Show your evaluator a cell you are going to draw when you have the slide ready under the microscope. Put your microscope away when you are finished.
>
> Time limit - 5 minutes
>
> Microscope - 5 marks
>
> Scientific Diagram - 10 marks

It should be noted that all of the teachers involved in this assessment had been experienced in the utility of the microscope and the design of the scientific diagram. A brief meeting amongst the examiners took place before the enactment of this particular performance assessment in order to discuss the relative scale of the scoring rubrics on the microscope task and the scientific diagram.

The biology teacher invited the other participating science teachers to critique her assessment tasks. The teachers collectively shared a satisfaction with the test design, and expressed appreciation over the specific items in the scoring rubrics. The teachers agreed that the delineation of the scoring rubrics should enhance the inter-rater reliability in this particular activity.

Results:

Task 1 = Microscope use = 5 marks total
Task 2 = Scientific Diagram of Onion Root = 10 marks total

Table 3A
Student 1 - Roger

Teacher	Task 1	Task 2
	Score /5	Score /10
Mr. P	5	6.5
Ms. A	4	6
Mr. H	5	7
Ms. K	5	6.5
Ms. J	5	6.5

Task 1 average = 4.8
Task 2 average = 6.5
Standard Deviation of Scores on Task One = 0.46
Standard Deviation of Scores on Task Two = 0.35

Table 3B
Student 2 - Anne

Teacher	Task 1	Task 2
	Score /5	Score /10
Mr. P	4.5	3
Ms. A	4.5	3
Mr. H	5	2
Ms. K	4.5	2.5
Ms. J	4.5	2

Task 1 average = 4.6
Task 2 average = 2.5
Standard Deviation of Scores on Task One = 0.22
Standard Deviation of Scores on Task Two = 0.50

Table 3C
Student 3 - Nancy

Teacher	Task 1	Task 2
	Score /5	Score /10
Mr. P	3.5	4
Ms. A	2.5	2
Mr. H	4	3
Ms. K	3.5	2
Ms. J	3.5	2

Task 1 average = 3.4
Task 2 average = 2.6
Standard Deviation of Scores on Task One = 0.55
Standard Deviation of Scores on Task Two = 0.89

Summary Of Inter-rater Reliability For Performance Assessment 3 - Microscope Skills

The Pearson's Product Moment Correlation Coefficient was used to calculate the inter-rater reliability between each of the examiners. This coefficient is representative of the strength of correlation of one examiner's score with another examiner's score. The range of possible positive correlations and their interpretations are as follows:

1.0 perfect positive correlation
0.8 very strong positive correlation
0.6 moderate to strong positive correlation
0.4 weak correlation
0.2 very weak correlation
0 probably no correlation (Fitz-Gibbon, Morris, 1978, p. 92)

Table 3R
Inter-rater Reliability Matrix of Five Raters in Performance Assessment 3 – Microscope Skills (Correlation Coefficients - Pearson's Product Moment)

	Mr. P	Ms. A	Mr. H	Ms. K	Ms. J
Mr. P		0.86	0.94	0.89	0.90
Ms. A			0.87	0.91	0.91
Mr. H				0.96	0.98
Ms. K					0.99
Ms. J					

Overall interrater reliability for performance assessment task 1 (Electronic Circuits) using Pearson's Product Moment Correlation Coefficient over 60 paired scores = 0.88

Summary Thoughts On Performance Assessment 3 - Microscope Skills

The third performance assessment activity implemented by our research group of science teachers yielded a very strong correlation of inter-rater reliability, r = 0.88. Sixty paired scores were utilized to calculate the overall inter-rater reliability based on Pearson's Product Moment Correlation Coefficient formula. These sixty paired scores produced a lower limit of 95% confidence interval of r = 0.83 and an upper limit of 95% confidence interval of r = .93.

This very positive result suggests very strong inter-rater reliability which is most likely a result of the very delineated scoring rubrics which clearly established exactly what skill was necessary in order to achieve the various marks. It was this design feature that our collective group of teachers felt was most responsible for the extremely strong inter-rater

reliability demonstrated in this activity. It is also probable that our research group had acquired more skill in the design and implementation of performance assessment tasks. On an interesting note, during the setup and implementation of this particular performance assessment activity, two members shared that they were about to implement their own performance assessment tests in their regular classes.

Teachers in this research project considered the face validity of performance assessment "Microscope Skill" as very strong because these tested items are part of the program of studies for Science 10 students in Alberta. The performance assessment tasks focused on the abilities that each student should be competent doing after completion of the microscope section in the science 10 program. One part of this performance assessment appeared to be quite feasible in a group setting, where a number of students could be completing the scientific diagram based on the images on the microscope, which would then be marked by the teacher at his/her convenience. The challenge is to observe the various students as they are actually manipulating the microscope. This aspect of the performance assessment process is difficult if the numbers of students exceeds four at a time, especially if there is only one teacher observing.

The group of science teachers working on this research project

were most definitely finding merit in the enactment of the performance assessment process. The general feeling in the research group was becoming more positive with each additional performance assessment enacted. As well, one must appreciate that this research project took place over a seven month interval, which is most certainly a long time for a group of busy science teachers, especially considering that there was no financial recompense.

Performance Assessment 4 - Uniform Motion

Measurement of the velocity of an air track vehicle was the theme for this set of performance assessment tasks. The student was supplied with an air track, timing apparatus, air vehicle, and starting gate. The air track vehicle was accelerated by using a rubber band launch mechanism and timed through the utility of electronic timing gates. Time allowed for the completion of this entire activity was five minutes. Each student was exposed to the problem and the scoring rubrics fifteen minutes before beginning the actual performance assessment tasks.

The scoring rubrics for this particular performance assessment involved:

- 1 mark for correct formula/relationship,
- 3 marks for appropriate apparatus use,
- 3 marks for correct answer and units.

The six teachers involved in the assessment of this performance activity were briefed into the logic and proper methods in the solution of the velocity of the air track vehicle. The three students chosen for this particular activity involved 2 grade 10 students and 1 grade 9 student. All three selected students had performed a similar experiment in regular science class using the same apparatus.

Results:

Table 4A
Student 1 – Betty

Teacher	Task 1	Task 2	Task 3	Total
	Score /1	Score /3	Score /3	Score /7
Mr. P	1	2.5	2.5	6
Ms. A	1	3	2	6
Mr. H	1	2.5	2.5	6
Ms. O	1	3	2.5	6.5
Ms. J	1	3	2	6
Ms. K	1	3	2	6

Average of total score /7 = 6.1

Standard Deviation of Scores on Task One	= 0
Standard Deviation of Scores on Task Two	= 0.26
Standard Deviation of Scores on Task Three	= 0.27
Standard Deviation of Scores on Total	= 0.20

Table 4B
Student 2 - Anita

Teacher	Task 1	Task 2	Task 3	Total
	Score /1	Score /3	Score /3	Score /7
Mr. P	.25	3	2.5	5.75
Ms. A	0	3	2	5
Mr. H	.5	3	2.5	6
Ms. O	0	3	2.5	5.5
Ms. J	0	3	2	5
Ms. K	0	2	2	4

Average of total score /7 = 5.2

Standard Deviation of Scores on Task One = 0.21
Standard Deviation of Scores on Task Two = 0.41
Standard Deviation of Scores on Task Three = 0.27
Standard Deviation of Scores on Total = 0.71

Table 4C
Student 3 - Jacob

Teacher	Task 1	Task 2	Task 3	Total
	Score /1	Score /3	Score /3	Score /7
Mr. P	0.25	3	1.5	4.75
Ms. A	0	3	1	4
Mr. H	0	3	2	5
Ms. O	0	3	1	4
Ms. J	0	3	2	5
Ms. K	0	3	3	6

Average of total score /7 = 4.8

Standard Deviation of Scores on Task One = 0.10
Standard Deviation of Scores on Task Two = 0
Standard Deviation of Scores on Task Three = 0.76
Standard Deviation of Scores on Total = 0.75

Summary Of Interrater Reliability For Performance Assessment 4 - Uniform Motion

The Pearson's Product Moment Correlation Coefficient was used to calculate the inter-rater reliability between each of the examiners. This coefficient is representative of the strength of correlation of one examiner's score with another examiner's score. The range of possible positive correlations are as follows:

1.0 perfect positive correlation
0.8 very strong positive correlation
0.6 moderate to strong positive correlation
0.4 weak correlation
0.2 very weak correlation
0 probably no correlation

Table 4R
Inter-rater Reliability Matrix of Six Raters – Activity 4 – Uniform Motion (Correlation Coefficients - Pearson's Product Moment)

	Mr. P	Ms. A	Mr. H	Ms. O	Ms. J	Ms. K
Mr. P		0.90	0.98	0.98	0.95	0.83
Ms. A			0.93	0.99	0.96	0.91
Mr. H				0.95	0.93	0.87
Ms. O					0.94	0.90
Ms. J						0.91
Ms. K						

Overall inter-rater reliability for performance assessment 4 (Uniform Motion) using Pearson's Product Moment Correlation Coefficient over 135 paired scores = 0.91

Summary Thoughts On Performance Assessment 4

Microscope Skills

This particular performance assessment activity garnered an extremely strong rate of inter-rater reliability of r = 0.91. The six science teachers that participated in this assessment activity were very pleased with this very strong correlation coefficient, which is extremely accurate due to the large number of paired scores which were utilized in the calculation of the "Pearson's Product Moment Correlation Coefficient." The upper limit of 95% confidence interval is r = 0.93 and the lower limit of 95% confidence interval is r = 0.88. Such a very high inter-rater reliability in the performance assessment process involving several adjudicators, in this case six, is of significance in the implementation of reliable performance assessments in the classroom.

The fact that this particular set of performance assessments utilized scoring rubrics that were actually demonstrated before the actual examination to each of the adjudicators, invariably contributed to the very high rate of agreement on scoring. The feedback from the teachers and students involved in the particular assessment activity were extremely positive coupled with an increased interest in the assimilation of such mode of evaluation into the regular evaluative context of the teacher's pedagogical situation.

As a researcher, one is especially confident of the result in this set

of performance tasks in lieu of the large number (135) of paired correlations used in the establishment of the overall correlation coefficient. By using a greater number of paired scores, one is able to reduce the error in the calculation of the "Pearson's Product Moment Correlation Coefficient."

The group consensus on the issue of face validity for this performance assessment activity "Uniform Motion" was that of strong validity based in the fact that this activity is a required laboratory experience for the grade 10 science student in the physics portion of the curriculum. In the class activity the students utilize a ticker tape timer device to calculate velocity; however, this device is an inconsistent timing device. The air track we used was considerably more precise with the digital timing equipment in comparison to the ticker tape timer. The digital timer is very accurate due to a precise digital timing mechanism with optical sensors which provided timing accuracy to +/- 0.001 seconds. Moreover, exposing students to more advanced technology in the science laboratory provides a closer view into the apparatus they will likely encounter in today's dynamic world of technology and science.

The science teachers involved in this research project felt that this particular assessment activity was well-suited to a science 10 classroom environment, where the teacher could rotate small groups of students through the activity and collect their written results. One science

teacher suggested that the students could actually videotape their performances so as to aid in the teacher's evaluation of their specific performances. The process of videotaping of the performances would also allow for future scrutiny that could be of value in the training and constant refinement of this labor intensive mode of assessment.

Feedback from the students during informal discussion after this performance assessment activity with the research science teachers was very positive. Students shared their willingness to participate in future performance assessments, which is an importance motivation for students to continue their studies in the sciences. If the perception exists that the study of science is dry and dull, educators must work to dispel such folly. Performance assessments demonstrate the promise to aid in the maintenance of adequate hands-on experiences which will undoubtedly aid to the students enjoyment and satisfaction in the study of science.

The trend emerging from these first four sets of performance assessments is that of increased inter-rater reliability for each successive task. Is is possible that our team is becoming more competent through practice? Is the design of our performance tasks improving? Our research group discussed these possibilities and felt overall that our skills were improving as we progressed. This would lead one to appreciate the role of effective instruction in the implementation of

performance assessments in a school district.

Performance Assessment 5 - Acid/Base Identification and Neutralization

This final performance assessment activity challenged the student to:

1) use the supplied apparatus and materials to determine whether the unknown solution is an acid or base;

2) describe to the examiners how to do this procedure;

3) neutralize this solution using the materials supplied and demonstrate checks for the neutral property of this solution.

The scoring for this set of performance assessment tasks was:

- description of procedure for determination of acid or base
(1 mark)

- hands-on technique in the determination of the acid or base
(2 marks)

- description and hands-on technique for neutralization of solution
(3 marks).

Each student involved in this performance assessment activity was given an opportunity to view these scoring rubrics individually, just before they initiated their performances.

It should be noted that one of the examiners in this activity was a substitute teacher, who had recently replaced one of the science teachers on a short term leave. This individual, Ms. S. was willing to join our research group at this late stage (final performance assessment activity). As the author of this study, I felt that her inclusion in the research group

would be of significant value in the domain of inter-rater reliability. The groups' initial hypothesis about the inclusion of this new untrained examiner was that she would likely not display the same degree of inter-rater reliability demonstrated between the experienced examiners.

The research group of science teachers met with the new addition to our team, Ms. S. and gave her a one hour introduction to our performance assessment project and then included her in this particular activity, "Acid/Base Identification and Neutralization." The examiners did review the scoring rubrics with the new assessor and quickly (ten minutes) overviewed the expected exemplars of performance at the various levels of competency. The fact that this particular instructor was not a science major may have affected her scoring of the performance tasks.

Results:

Table 5A
Student 1 - Jan

Teacher	Task 1	Task 2	Task 3	Total
	Score /1	Score /2	Score /3	Score /6
Mr. P	1	1.5	3	5.5
Ms. L	1	1	3	5
Mr. H	1	1	3	5
Ms. J	1	1.5	3	5.5
Ms. S	0.5	2	2	4.5

Average of total score /6 = 5.1
Standard Deviation of Scores on Task One = 0.22
Standard Deviation of Scores on Task Two = 0.42
Standard Deviation of Scores on Task Three = 0.45
Standard Deviation of Scores on Total = 0.42

Table 5B
Student 2 - Trent

Teacher	Task 1	Task 2	Task 3	Total
	Score / 1	Score /2	Score /3	Score /6
Mr. P	1	2	2.5	5.5
Ms. L	1	2	3	6
Mr. H	1	2	2.5	5.5
Ms. J	1	2	3	6
Ms. S	1	2	2	5

Average of total score /6 = 5.6
Standard Deviation of Scores on Task One = 0
Standard Deviation of Scores on Task Two = 0
Standard Deviation of Scores on Task Three = 0.42
Standard Deviation of Scores on Total = 0.42

Table 5C
Student 3 - Ross

Teacher	Task 1	Task 2	Task 3	Total
	Score / 1	Score /2	Score /3	Score /6
Mr. P	1	2	3	6
Ms. L	1	2	3	6
Mr. H	1	2	3	6
Ms. J	1	2	3	6
Ms. S	1	2	2	5

Average of total score /6 = 5.8
Standard Deviation of Scores on Task One = 0
Standard Deviation of Scores on Task Two = 0
Standard Deviation of Scores on Task Three = 0.42
Standard Deviation of Scores on Total = 0.45

Table 5D
Student 4 - Jim

Teacher	Task 1	Task 2	Task 3	Total
	Score /1	Score /2	Score /3	Score /6
Mr. P	1	2	2	5
Ms. L	1	2	1	4
Mr. H	1	2	2.5	5.5
Ms. J	1	2	1.5	4.5
Ms. S	1	2	1	4

Average of total score /6 = 4.6
Standard Deviation of Scores on Task One = 0
Standard Deviation of Scores on Task Two = 0
Standard Deviation of Scores on Task Three = 0.65
Standard Deviation of Scores on Total = 0.65

Summary Of Inter-rater Reliability For Performance Assessment 5 - Acid/Base Identification and Neutralization

The Pearson's Product Moment Correlation Coefficient was used to calculate the inter-rater reliability between each of the examiners. This coefficient is representative of the strength of correlation of one examiner's score with another examiner's score. The range of possible positive correlations and their interpretations are as follows:

1.0 perfect positive correlation
0.8 very strong positive correlation
0.6 moderate to strong positive correlation
0.4 weak correlation
0.2 very weak correlation
0 probably no correlation

Table 5R
Inter-rater Reliability Matrix of Five Raters for Activity 5
(Correlation Coefficients - Pearson's Product Moment)

	Mr. P	Ms. L	Mr. H	Ms. J	Ms. S
Mr. P		0.91	0.97	0.97	0.75
Ms. L			0.85	0.98	0.74
Mr. H				0.89	0.59
Ms. J					0.79
Ms. S					

Overall inter-rater reliability for performance assessment task 5 using Pearson's Product Moment Correlation Coefficient over 120 paired scores = 0.83

Summary Thoughts On Performance Assessment 5 - Acid/Base Identification and Neutralization

This final group performance assessment activity achieved a high degree of inter-rater reliability of r = 0.83. It should be noted that one member of the science teachers involved in the judging of this activity was new to the research group. This was also the first time in which this particular teacher was involved in this performance assessment research project. As one reviews the matrix comparing the inter-rater reliability between the various adjudicators, one does indeed notice that one of the evaluators does not generally show the degree of agreement as is evidenced by the other adjudicators (r = 0.59 to r = 0.79). If you view the matrix comparing the reliability coefficient from teacher to teacher, you quickly realize that Ms. S is consistently lower in correlation coefficient as compared to the other examiners.

In reviewing the certainty in the calculation for the overall reliability of this performance assessment activity, which was $r = 0.83$, the upper limit of 95% confidence interval is $r = 0.88$ and the lower limit of 95% confidence interval is $r = 0.77$. These figures reflect a strong positive inter-rater reliability amongst the adjudicators. The fact that this fifth performance assessment activity yielded a slightly lower inter-rater reliability $r = 83$ than the last activity $r = 0.91$ is due to the inclusion of a "rookie" examiner, who was new to the performance assessment process.

Face Validity of this performance activity was seen as strong by the participating group of science teachers. The fact that the theme of this activity is covered by curriculum in the science 9 and 10 program of studies (Alberta), reinforces its validity and necessity in the evaluative arsenal of the secondary education science teacher. It should be noted that it was unanimously agreed that the performance assessment process utilized by our group of science teachers was seen as being authentic, for the student not only had the responsibility of relaying the theory associated with the task, but also had to provide actual hands-on performance to reinforce his/her theories. This multi-sensoral approach was viewed as an important process in the education of the science student. Virtually all teachers in this research group agreed that students should be able to demonstrate their skills in an authentic

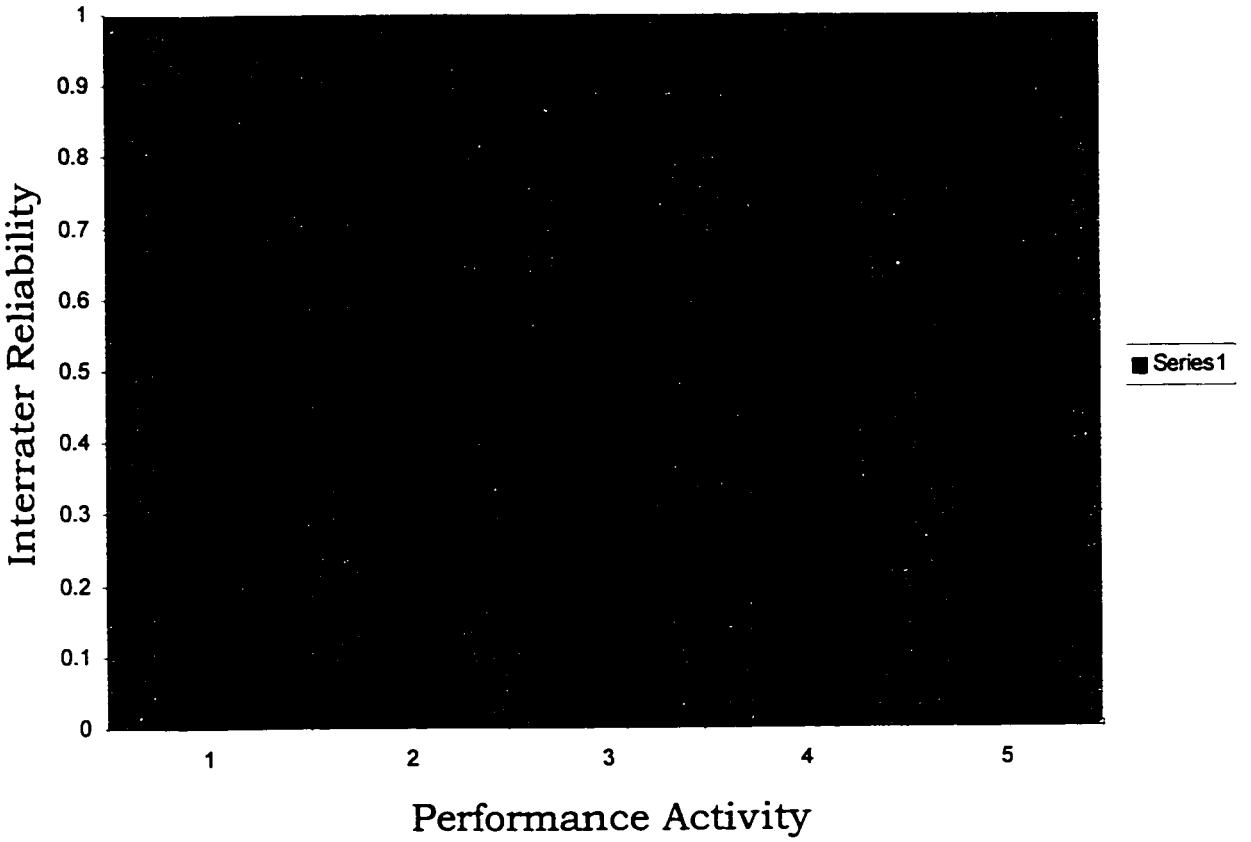
hands-on manner, in addition to the traditional pencil and paper examination paradigm that dominates in our evaluative schemata.

Chapter 7

Summary of Performance Assessment Tasks Designed and Implemented By Collaborative Research Group

This collaborative research project involved a group of eight secondary science teachers at an urban composite high school. These teachers cooperatively developed and tested a variety of performance assessment tasks for the duration of one school year, 1995 to 1996. The development and testing of each performance assessment task was done as a democratic group effort, with the researcher acting as an equal partner with the other 7 science teachers. Analysis of the performance assessment results included standard deviation, Pearson's Product Moment Correlation Coefficient, and face validity evaluation. The inter-rater reliability achieved by this research project varied from a low of $r = 0.83$ to a high of $r = 0.91$. These results indicate very strong inter-rater reliabilities, which reinforces the research of Gipps (1994) that found that with the inclusion of "clear rubrics and training for markers, and exemplars of performance at each point or grade, levels of IRR (inter-judge reliabilities) can be high" (Gipps, 1994, p. 104).

Interrater Reliability Comparison Graph



The inter-rater reliabilities demonstrated by the examiners in this study displayed a constant improvement from the first set of performance assessment tasks to the fourth set of performance assessment tasks (IRR= 0.83, 0.87, 0.88, 0.91 respectively). Yet the fifth set of performance assessment tasks showed a decrease in inter-rater reliability following the forth set from 0.91 to 0.83. This decreased inter-rater reliability in the fifth set of performance assessment tasks was due to the fact that one of the examiners in this particular set of performance tasks was new to this research project, and had not participated in the other four

performance assessment tasks. When one reviews the inter-rater reliability matrix comparing each of the examiner's sets of scores with each other, one notices that one examiner is not nearly as consistent as the other four examiners (see chart below). It so happens that this new examiner, Ms. S. was indeed the one who displayed the weakest correlation coefficient with the other examiners. This leads one to appreciate the importance of training and experience in the delivery of performance assessment tasks.

Table 5R
Inter-rater Reliability Matrix of Five Raters in Performance Assessment 5
(Pearson's Product Moment Correlation Coefficient)

	Mr. P	Ms. L	Mr. H	Ms. J	Ms. S
Mr. P		0.91	0.97	0.97	0.75
Ms. L			0.85	0.98	0.74
Mr. H				0.89	0.59
Ms. J					0.79
Ms. S					

As seen in the preceeding correlation matrix, examiner (Ms. S.) did not display a high degree of agreement with the others. In examining the scores from Ms. S, one finds that she was consistently marking the students lower than the other examiners. It would appear that Ms. S. the new teacher to this project did not operate on the same scale of adjudication as the other four examiners. Is it possible that the scoring rubrics were too general for this activity? Did we not clearly discuss with our new recruit, the levels of achievement associated with each grade assigned? Of interest too, is the fact that Ms. S. was the only teacher

who was not a science teacher by training. Her major was in another area, yet her assessments would have had closer correlation with the others simply by adding a small amount to each of her assessments. She was operating on a more critical level than the rest of the team. Is it possible that our assessment team became more liberal in assigning grades for various performance tasks as the study weathered on? These could form the nucleus of issues for future researchers.

The proceeding charts summarize the inter-rater reliability correlation coefficients for each of the performance assessment activities in this project. Pearson's Product Moment Correlation Coefficient was used for the formulation of these results. This coefficient is representative of the strength of correlation of one examiner's score with another examiner's score. The range of possible positive correlations and their interpretations are as follows:

1.0 perfect positive correlation
0.8 very strong positive correlation
0.6 moderate to strong positive correlation
0.4 weak correlation
0.2 very weak correlation
0 probably no correlation

(Fitz-Gibbon, Morris, 1978, p. 92)

Summary Of Inter-rater Reliability For Performance Assessment 1

Electronic Circuits

Table 1R
Inter-rater Reliability Matrix of Five Raters (Correlation Coefficients - Pearson's Product Moment)

	Mr. P	Ms. A	Mr. D	Ms. K	Mr. A
Mr. P		0.95	0.81	0.86	0.92
Ms. A			0.94	0.95	0.91
Mr. D				0.96	0.81
Ms. K					0.92
Mr. A					

Overall inter-rater reliability for performance assessment task 1 (Electronic Circuits) using Pearson's Product Moment Correlation Coefficient over 60 paired scores = 0.83

Summary Of Inter-rater Reliability For Performance Assessment 2

Measuring Density

Table 2R
Inter-rater Reliability Matrix of Seven Raters (Correlation Coefficients - Pearson's Product Moment)

	Ms. O	Ms. J	Ms. A	Mr. H	Mr. D	Ms. K	Mr. A
Ms. O		0.98	0.98	0.92	0.91	0.97	0.96
Ms. J			0.96	0.98	0.94	0.98	0.98
Ms. A				0.94	0.84	0.91	0.91
Mr. H					0.91	0.92	0.94
Mr. D						0.98	0.98
Ms. K							0.99
Mr. A							

Overall inter-rater reliability for performance assessment task 2 (Measuring Density) using Pearson's Product Moment Correlation Coefficient over 78 paired scores = 0.87

Summary Of Inter-rater Reliability For Performance Assessment 3

Microscope Skills

Table 3R
Inter-rater Reliability Matrix of Five Raters (Correlation Coefficients - Pearson's Product Moment)

	Mr. P	Ms. A	Mr. H	Ms. K	Ms. J
Mr. P		0.86	0.94	0.89	0.90
Ms. A			0.87	0.91	0.91
Mr. H				0.96	0.98
Ms. K					0.99
Ms. J					

Overall inter-rater reliability for performance assessment task 1 (Electronic Circuits) using Pearson's Product Moment Correlation Coefficient over 60 paired scores = 0.88

Summary Of Inter-rater Reliability For Performance Assessment 4

Uniform Motion

Table 4R
Inter-rater Reliability Matrix of Six Raters (Correlation Coefficients - Pearson's Product Moment)

	Mr. P	Ms. A	Mr. H	Ms. O	Ms. J	Ms. K
Mr. P		0.96	0.98	0.98	0.95	0.83
Ms. A			0.93	0.99	0.96	0.91
Mr. H				0.95	0.93	0.87
Ms. O					0.94	0.90
Ms. J						0.91
Ms. K						

Overall inter-rater reliability for performance assessment 4 (Uniform Motion) using Pearson's Product Moment Correlation Coefficient over 135 paired scores = 0.91

Summary Of Inter-rater Reliability For Performance Assessment 5

Acid/Base Identification and Neutralization

Table 5R
Inter-rater Reliability Matrix of Five Raters (Correlation Coefficients - Pearson's Product Moment)

	Mr. P	Ms. L	Mr. H	Ms. J	Ms. S
Mr. P		0.91	0.97	0.97	0.75
Ms. L			0.85	0.98	0.74
Mr. H				0.89	0.59
Ms. J					0.79
Ms. S					

Overall inter-rater reliability for performance assessment task 5 (Acid/Base Identification and Neutralization) using Pearson's Product Moment Correlation Coefficient over 120 paired scores = 0.83

Each of the five performance assessment tasks utilized by the research group of science teachers yielded very strong inter-rater reliability correlation coefficients right from the first performance assessment task to the last. The preceding charts clearly show how each examiner's scores compared with each other. The only performance assessment task which showed slightly less agreement amongst examiners was the last activity, acid/base identification and neutralization. The examiner who displayed a lower degree of agreement on her scores was indeed the new teacher to the research group. She was a substitute for one of the science teachers, and not a science major herself. This result has important ramifications for the training of examiners in the performance assessment process. The examiners must

have agreement in the levels of performance for each assigned grade before the performance tasks are implemented.

The strong positive correlation between examiner's scores in this research project indicate that performance assessment tasks are a reliable mode of assessment in science, especially where scoring rubrics are utilized, as in these five performance assessment activities.

It was unanimously agreed, by all teachers involved in this project, that the performance assessment tasks created and tested were indeed valid. This collaborative view was based on the performance tasks inclusion in the curriculum of science studies as suggested experimental activities and/or skills.

An additional benefit of this project was the ability to gain insights into each student's constructs of the particular theory/skill being tested. Students did not necessarily achieve results on their performance assessment tasks with the same degree of success or failure as on their written exams. This would suggest that science teachers are well advised to incorporate performance assessment tasks as a means to better assess the overall abilities of each student. Moreover, the students themselves were unanimous in their appreciation of the performance assessment process. In every instance, students gave positive feedback on the particular performance assessment task they

were involved in. The hands-on approach was viewed as an enjoyable addition to the traditional pencil-and-paper exam.

The teachers involved in this research project did express a concern for the time requirements necessary for the development and implementation of the performance assessment activities. It was generally agreed by the participating science teachers that the performance assessment process is a time-demanding process, requiring teachers willing to put in the extra effort to set-up equipment, etc. which is no small task in the ever-demanding schedule of the everyday science teacher. The cut-backs in teacher preparation time were seen as limiting factors in the development and integration of performance assessment tasks in the science classroom. Perhaps a future study on the time issue of performance assessment development and delivery would address these concerns shared by the participant teachers.

Aside from the time requirements, the participant science teachers were in complete agreement over the value of incorporating performance assessments in the science classroom. In fact, most teachers developed and integrated performance assessment tasks into their own particular pedagogical situation in the following year to this study. This was quite a considerable turn-around from the beginning of this study in which many of the participating science teachers shared that performance assessment activities were not that feasible in their own class.

This project has clearly demonstrated that performance assessment tasks can be a reliable and fruitful means of assessing the skills of our science students. These results must be tempered by the fact that this was a collaborative effort with eight teachers. Moreover, the performance assessment tasks involved a very small sample of students who volunteered their services at the end of the school day. This most certainly does not parallel a class of thirty grade nine science students with one teacher. I feel that the next logical step from this project would be for continued research into the enactment of performance assessments in classroom settings with various numbers of students, from twenty to thirty. One teacher would also be present during these performance tasks, which is consistent with the realities in today's schools.

In the year following this project, 1996-1997, five of the teachers involved in this collaborative project designed and implemented performance assessment tasks in their own classrooms during regular hours. I was able to share some time with each of these teachers on reflection about the merits of their efforts. With all five teachers, one of which was myself, we found that the performance assessment activities were a valuable addition to the assessment paradigm in our classroom. Four of the teachers shared a desire to continue the development and implementation of performance tasks in their science classes, but the

fifth teacher did not enjoy the time demands that the performance assessment process involved in apparatus set-up and task completion. With all five instructors that integrated performance assessment tasks in the following year, the students completing the tasks recorded their results and handed them in for grading at a later date in the majority of their performance tasks. Three of these teachers did include performance assessment tasks which involved immediate one-on-one judging, but the classes where this occurred were generally quite small, from fifteen to twenty students.

If one is to judge this project a success, then one would expect to see an increase in the implementation of performance assessment tasks at the school involved in this study. In this regard, performance assessment tasks are indeed increasing in design and delivery at the test school. Another benefit worth mentioning, is the result of the collaborative research with the entire science department at the test site. The collaborative spirit has continued beyond the duration of this study (one year) and fostered a cohesive and cooperative science department very willing to assist and share our instructional resources. This is no small feat in the infrastructure of any school.

Chapter 8

Research Group Reflections

Upon completion of the five sets of performance assessment tasks, the group of eight science teachers involved through the duration of this research project shared their final thoughts on the strengths and weaknesses of the performance assessment process and its place in the evaluation schemata of the teacher. These collaborative researchers were also asked to comment on the validity of the performance assessment tasks utilized in this research project. These reflections were done verbally with the assistance of an audio tape recorder. I have transcribed the audio tapes exactly as they were recorded.

Teacher 1 – Mr. A.

"I think it's (performance assessment task) a very good way of determining what students really know about the subject area they are studying. It's one thing to get a theoretical background but quite a different thing to explain it through activities that relate to the theory that they are studying."

"I think that this is a valid assessment tool. I feel that it measures what we are wanting it to measure. The only drawback that I do see is that it would take a great deal of time to measure all the variables and concepts that we want to measure and the perception of one individual, one teacher without another may be quite different than working with a

group of teachers. As well, some students may want to and actually be able to do the activity using materials other than the ones before them."

"I think that my attitude was a little negative toward it (performance assessment) at first, but I think that I am now encouraged by performance assessment and the measuring of it in a more consistent way than we had ever done before."

"I think that the student's perception of what he thinks he understands and what he actually does comprehend depends on doing the experiments (performance assessment tasks). Things become real to him. He is actually working in a real environment. Then he can try to equate the two, the theory with actual practice, and see if his assertions are valid. I think from these things (performance assessments) other concepts will grow. There are many off-shoots to these performance assessments which the students are doing. I was really amazed at the way some of these students went about to find density. The steps that they took, in some instances, really surprised me and some of these alternative methods were valid."

Teacher 2 – Mr. H.

"I liked the fact that it (performance assessment) made the kids more conscious of the stuff they were doing. I found it very difficult to mark them in a fair manner. It's not something I would lean toward very much because the kids can watch other kids do some of the stuff before I can watch them, so that's not really fair."

"There's too many of them to watch, unless like my group of Chemistry 30 which is only around ten students, but with a group of twenty five or thirty students, you couldn't possibly evaluate them all in the same class. If you do some of the students before others, it's not real fair to them, because some students will be able to see other students perform the particular activity before others. But, just the fact that your watching them do activities makes them (students) more conscious of their work. I received better lab results than I normally attain and I got more students involved in the activities. Previously, some of the students would sit around and watch others do everything. I think from that point of view it's definitely useful and with some of the things I do in class, I will do that (performance assessment tasks) again with the students so that they will know that I am watching and checking off things off but in terms of weighting it into the marks, it's not going to count for a whole lot."

"In the beginning (of the research project) I thought that this process would be impossible to do with a class of students, and now I know its not impossible, but it is very difficult. I'll try to make use of it more than I have, because it does help with the students' perceptions of how they do their labs and that sort of thing, so that's a definite change. Before, I thought that there's no way I can possibly do this stuff (performance assessment tasks) with all the students I have. Teaching

four courses all day long, it's just hectic! There's enough stuff to try and do without having to do this stuff, too."

"This thing (performance assessment task) with the uniform motion air track was obviously a thing that the students enjoyed working with! You can use this thing easily and get good results from it. So I thought that one worked out pretty well."

"The microscope one (performance assessment task) I found difficult to do, because there is no way I watch them all (a large class) and there were so many things on the checklist that we had to cover that it made it very difficult to do that one."

"The acid/base one worked out very well, because I thought we covered the things we needed to cover. It was a short enough assessment that you could watch them fairly easily. The thing that Ms. O did with the lab, she said that the lab exam with kids is pretty much impossible unless you have an assistant that can set up a lot of the things for you, so that you only have to worry about the evaluation. If you have stuff set up and you have to reset it up when the next kid comes in, the time constraints are just too much. I don't see how you can do that with a class of thirty kids unless you want to spend time outside of class time doing it such as the kids come in at noon hour, but this takes excess time and we are too constrained."

"I enjoyed doing this research project on performance assessment with you."

Teacher 3 – Ms. J.

"I've always really liked the idea of performance assessments and how useful they can be. It really tests what we are trying to do, not just having students write out theory, but actually doing the activities."

"The strengths of performance assessments as I see it, are that we are getting out of the kids what we do in class. If you are trying to teach a student how to do a certain thing, you want them to see the performance, not simply explain it on paper. I'm more concerned that kids can actually perform, not just memorize because I find that a lot of students have problems explaining, whereas if you give them something to do, they may be able to do it. Sometimes they are given a lower mark because they are not able to explain yet they still know what's going on and this (performance assessment) shows that they do understand what's going on and this is what real life is all about."

"The time factor is a challenge with performance assessments. With a large class with a number of students this could be difficult but with a smaller class it might work out for you. Usually with a bigger class it (performance assessments) takes up much time and the way things are already, we don't have a lot of time. That's my biggest concern, time!"

"The other concern would be as an individual teacher and the amount of time we have outside of school; its really taken up. It would be difficult to find the time to develop the tasks and the associated scoring

rubrics. The evaluation of the performance assessments is not too bad, but the development of the tasks and scoring rubrics is quite time-consuming, especially for a beginning teacher, as I am with the sciences. It would take up a whole lot of time and there just isn't that time around!"

"All of the performance assessment tasks which we did were very valid. The teachers that made up each task knew exactly what they wanted. In some things teachers expect students to do things one way, or they want to see understanding."

"I see this process (performance assessment) as very useful and important. We should take the time to do more of these if possible."

Teacher 4 – Ms. O.

"The strengths that I see with performance assessments would be associated with a good set of scoring rubrics. I think that you could usually come up with a very reliable and repeatable marking system. If the rubrics are very general and not detailed, then we started getting more variance in the marking aspect of it. One of the things that I would like to see is part of our marks in a lab-oriented course based on performance assessments. We do too little of that! The kind of testing that we do in chemistry, we mark the lab and the report that the students do, but it doesn't necessarily test their ability to manipulate the equipment itself, to recognize the proper equipment, and evaluate this."

"One of the weaknesses that I see and one of the things that we would have to overcome would be the feasibility with large groups of students. I think it would have to be set up in a lab exam manner and this would be quite time-consuming. I don't see myself being able to take a class of twenty-eight to thirty students and be able to observe all of their actions as they are doing the lab in one given period. We could either do it group by group; but, there would be questions of fairness. Some labs are more difficult to perform, and some procedures are more complex than others, like titration is a more involved process than filtration, or something like that."

"In terms of validity I thought that these (performance assessments) were really neat especially for the general science class. I would like to see us doing more of these (performance assessments). What I liked about the way that we did it is that we put large amounts of equipment there and just allowed the student to use which ever method they wanted to, like the acid/base one (performance assessment task). I appreciated that one, maybe because it dealt with chemistry a little bit more than the others. There is more than one solution and students have the flexibility of choosing a technique that shows us to an extent, different thinking patterns in the students. I don't think that we should necessarily treat all students exactly the same way, because they don't all learn in the same way. One process might not be better than another one, it's just different. In real life what you and I do to arrive at a

solution to a problem differs. We might take two different roads and still both arrive at an equitable and proper answer. I enjoyed that acid/base testing activity for that purpose."

"One of the things that I notice is that we kind of learned as we went along. We as teachers would need some kind of inservice or training. The more we do that, the better we get at it. By the time we had done four or five performance assessments, I was thinking some of the students are going to do this and some of the others something else. So my scoring rubrics were very specific. They (students) either did it or they didn't. I think as we practiced, like everything else we would get better at performance assessments."

"Now, how would we do performance assessments on a government exam? That would be interesting. I would like to see that, like on a diploma exam such as in physics and chemistry where they have a written component and a performance assessment component. I could see problem there. Number one, would all of the schools have the equipment, for example some of the smaller high schools? Would they send around a group of staff to do this? If they did this students could talk from one school to another, which would be a problem. They would have to train teachers so that reliability for all students would occur. The grading would have to be at the same playing field or level. That would be interesting. I don't know if that is coming in the future."

"Performance assessment was a new word to me before this project. I had not read anything on it and now I do know about it. It's something that I want to incorporate in my chemistry 20 class, for example. I was thinking the other day how I could do something along these lines with math. I haven't been doing manipulatives in math and this would be a way to test their physical understanding, fractions for example or theoretical things like equations solving or algebraic expressions. If you could put it in terms of manipulatives so that they (students) can do it. I know they do exist, but I haven't been using them. I might be interested in expanding along this line."

"I would like to see that if Alberta Education intends to go in this direction (performance assessments), that we will get training in making a rubric and also develop specific assessments for each course. I would like to see that!"

"I think that reliability would increase with time and practice. Even compared to multiple choice or written exams, they (performance assessments) could be reliable. Teachers will probably model their teaching based on what the government exams will look like. Their students will probably have a better chance at getting a higher mark. If a teacher does not concentrate on exam content and style this could affect exam results. Does this mean that students of teachers who teach to the exam know more about chemistry and physics then those students of

teachers who do not? The government exams do not cover all of the curriculum, they only cover a portion of it."

Teacher 5 – Ms. P.

"Students that blow written exams may have the opportunity to show their potential in performance exams. I think that this is the main reason for using them (performance assessment tasks). This relates to the hands-on types of kids."

"Unfortunately, I still do not think that this is possible to do in a class of twenty-five or greater with regular constraints unless you were given somebody to help you out. If you team taught or something similar it might work out. I still haven't figured out a way logically that you can do it (performance assessment tasks) as a class. I had a hard time seeing what three people were doing. I could see one, but beyond that I am not focused enough. Remember when we were doing the biology performance assessment? I had to ask other teachers if the students had completing this or that. Hard to do! Really hard to do! I would need a lot of preparation time in order to do this."

"I think that these tests were very valid. We were very consistent amongst ourselves. In a class with only one teacher evaluating, the teacher needs to be consistent with all his kids. A teacher would be (reliable) if they had set their criteria and knew what they are looking for, unless they are in a really bad mood one day. I think this is a very valid assessment process. I thought that the experiments we did (performance

assessments) were really good, the way they were set up and everything. It's just that incorporating them into a real life classroom is a tough one."

"Now I see performance assessments as more feasible since working on this than it was before. A simple activity where you set the kids up, you've got to have the right room for it, and maybe have kids coming in at noon or after school. Now it's starting to become more realistic to me (performance assessments). It still isn't totally realistic, but it's getting there!"

"I think it will be feasible with tiny classes, like science 14 or the I.O.P. students, maybe I'll try it. With my science nines, it would not be feasible because they were quite a handful, especially where experiments were concerned! Maybe I could do it at noon or after school. But I would have to give them class time out and I don't think that the administration would go for that too much. If they came in at lunch to do their performance assessment, could I let them out early in class? You've got to give them that sort of thing, or why else would they come in at noon."
"It was quite neat to do this research."

Teacher 6 – Ms. A.

"I think that it is a really good idea to look at performance assessment, especially for the lab tasks. People are supposed to be able to perform in high school science and I support the performance assessment idea. Initially I supported the performance assessment idea, and I still think it's a great idea, so I'm glad to see that we're looking at it.

I think that what we did at the school here showed that it's good for people to collaborate in order to share ideas and get some varied perspectives and backgrounds on performance assessment. I guess I already knew about performance assessments and what it was about. It seemed like a fairly easy type of task to do and setup. So my perspective really hasn't changed much. Initially I supported the idea. I think it's a good way to see what students can achieve and it holds true!"

"For validity, when you setup a performance assessment task, you have to be very specific and by being more specific you make it more valid. This way you are testing what is easily recognizable and that adds to the validity. When you setup your performance assessment task, it must be specific in order to measure more easily."

"As I said earlier, I like the performance assessment idea so I don't think that I've changed. I like the way that we collaborated on this project. I think it's a good way to generate ideas and also get a look at what other people can do in their disciplines such as chemistry, physics, the more specialized areas."

"I think this was a good start. We dappled in each of the disciplines and learned how we could apply performance assessment. This is worthwhile pursuing in further depth to continue some collaboration and to set up some performance assessment tasks that we could actually incorporate into our science 10 or whatever courses that we teach. Hopefully, it can build from here. This is a good start."

"Here are some thoughts on actually doing a performance assessment in a science lab or classroom. To do a performance assessment you have to be really specific in the task that you are asking the student to do. You have to schedule the task somehow, so that you can individually deal with the student or deal with two or three at a time. That is really difficult to do when you have a class of x number of students. So it does take considerable planning to schedule a performance assessment task. If you are short of time or if you have a large class, it's definitely difficult to do."

"When we went through these tasks, we discovered lots of little loopholes that you fall into like resetting the equipment, lack of materials, alternative way of solution that you haven't thought of. How do you evaluate those? There's many little things that crop up that you cannot necessarily predict, but you have to deal with it and maybe change your performance assessment task once you've done one or two trials. Try them out and hopefully perfect them with time. They (performance assessment tasks) definitely involve more commitment from the teacher, more time and more planning. Those are some things to consider as well."

Teacher 7 – Ms. K.

" I felt that this research project was very good at getting the school more involved in assessment and changed my perceptions of what I thought performance assessment was. I initially thought it was just

doing hands-on activities and evaluating just on what you could see students doing. Now I realize that it also includes things that the students can write, record, and ask questions based on what they have done. I guess I have already been doing quite a bit of this (performance assessments). I like to see evaluation of the skills. Unfortunately, if you have a very large class then you really have to be very particular on looking at very specific or few skills in order to really get a good analysis of what students are able to do."

"I think that this method is valid as long as your quite specific in what you're looking for and not leaving it vague. I think we noted that when we started some of the descriptors as to what was a skill was, had different interpretations. If it is spelled out quite clearly what is being looked for and what constitutes a particular mark, then it's quite valid. It has to be quite specific."

"The scoring rubrics was something that I hadn't dealt with before. The way that Alberta Education is pushing for more of this to occur, they should include more teacher resources and examples, such as what we have done, so as to give teachers guidance as to what is to be done. This is not that difficult to do, if you know what you are doing."

"I think that this was a good experience for us to get together and share ideas on different courses. It encouraged us in this assessment mode, and it does not have to be that difficult to do performance assessments if you use the proper means."

Teacher 8 – Mr. J.

"I certainly have a better appreciation of the rubrics and how similar teachers will evaluate based on observations. Yet, there are many differences as to what is competency and what is not. I think of the performance assessment where the boys were measuring density. Some teachers would look at the performance in a similar student and grade what was thought appropriate, yet there was still a varying in the results. I tended to be personally more lenient in my marking; whereas others, who were more attuned with the procedure, were more critical."

"When we did the experiments in the biology lab with the microscope, I found that there was too many criteria to evaluate. Too complex to really do a good job. I didn't think with that criteria, we could do a really good job on that. From my own experience in this type of evaluation, I use a smaller scale from one to five. Rarely do you use five and most often you're using the first four."

" I certainly have more appreciation of it (performance assessment tasks) now. I've used it for three years in the classroom doing various evaluations in the workplace and off-site evaluations. I find it the most flexible tool to evaluate students on a similar basis at different work-sites. You need some sort of continuity of measurement when you are evaluating. This method (performance assessment) seems to be the best, personally. I think the students feel comfortable with this form of evaluation."

"I think these tests measure the performance that is displayed. There is no question at either end of the scale. No one absolute form of evaluation is always effective. When you start to grade performance assessments of those who are reasonably competent, that the type of scale will show a range of success.

"Certainly, I think it's (performance assessments) valid in my own case with my own evaluations. When I evaluate a student who is successful, I can validate that measure of success by watching them succeed in whatever they are doing outside of the school. That validates what we have already seen in school. In the pure science lab, I am not sure how you could transfer those skills as being valid competencies. Where would you measure that in a work site? Maybe in another learning environment. That's what validity is to me. We can measure them as competent; but, if they (students) can move on and show competency in another environment, that should validate our evaluations on their performance."

"I think the biggest positive for me is to see other teachers using it (performance assessments). As I said this isn't new stuff for me, but I haven't done too much of it in a classroom. For those teachers who were involved in it (performance assessment research project) I know that it was ground breaking stuff. It was new methods of evaluating. I think that anything that causes teachers to stop and question some of our sacred cows, such as formal evaluation, the typical written exam . . . .

anything that moves us in a direction where we start to question our traditional methods. I think that this process is good!"

"I think that implementation of this type of product for the classroom is going to take time. Teachers will have to rethink what they are teaching in the classroom in order to properly evaluate performance assessment tasks. This is good! I believe it is always healthy for teachers to question what it is that they are doing in the classroom. What is their chosen method of evaluation? Is it effective? If performance assessments lead to insights in the evaluation gaps of a teacher, then it most definitely has value. I think that this is valid. To change methods just for the sake of change, no! I don't think too many people would do this. I don't prescribe to this either. There has to be a want and desire on the part of the teacher to evaluate their own effectiveness and what it is that they hope to accomplish in the classroom. If they can find success in their particular classroom with this type of assessment, great!"

Summary of Teacher Reflections

Certain themes were consistent in the reflections of the teachers who participated in this year-long research project. The issue of time was very pronounced from virtually every participant. The reality of decreased preparation time coupled with increased class size has minimized time available for the development and implementation of performance assessment tasks in the science classroom. Each of the

teachers in this project shared a strong concern for the time necessary in this mode of assessment; yet, they appreciated the collaborative process as an agent to assist in this process.

Class size was consistently seen as a limiting factor on the implementation of performance assessment tasks in the classroom. Performance assessment was seen as feasible in smaller classes, but not realistic in larger classes. It would be interesting, from a research perspective, to attempt performance tasks in a large class, of over thirty students, and analyze the results.

Every single teacher saw the performance assessment tasks as valid. These comments were coupled with the apparent necessity to make science more hands-on in order to maintain interest level in the students. Performance assessment was seen as a bridge between theory and practice. Many of the participating science teachers commented on the role of specific scoring rubrics in enhancing the validity and reliability of the performance assessment process. One teacher put it quite succinctly, when she mentioned that specific scoring rubrics help ensure that one is indeed evaluating what one thinks he/she is evaluating.

This collaborative research project generally left a positive and cohesive attitude amongst its participants. Each of the participating teachers shared a desire to continue this mode of assessment. Moreover, the collegial sharing enhanced the rapport between the teachers in this

department. This rapport has continued over the following two years since the collaborative research project ended.

The teachers generally spoke of enjoying this study, which is quite surprising given their busy schedules coupled with the fact that they were not given any financial remuneration. The attitude of the teachers throughout this project was very positive and cooperative. I was initially concerned that interest in this research project would reduce over the period of one year, but much to my amazement, the eight participating science teachers maintained their enthusiasm and commitment. Perhaps, the simple gathering as a group facilitated some of the positive feelings, which resulted from this study. I personally found this process very rewarding and invigorating.

The importance of clear and specific scoring rubrics was shared by each of the teachers. It was viewed that effective rubrics would enhance the examiners' abilities to maintain consistency from student to student. Rehearsal of the performance assessment coupled with clear delineation of each performance level was seen as a critical ingredient in success performance tasks.

This research project has proven to be a rewarding process for myself as the teacher/researcher. A strong collaborative spirit continued on in this science department following this project. I firmly believe that the sharing of knowledge empowers the human spirit. Performance assessment allows that spirit to fly.

Bibliography

Andersen, H., and Koutnik, P., (1972) Toward More Effective Science Instruction in Secondary Education. The Macmillan Company, New York.

Ary, A., Jacobs, L., and Razavieh, A., (1990) Introduction to Research in Education. (Fourth Edition), Holt, Rinehart and Winston, Inc., Fort Worth, Texas.

Ashworth, A., (1982) Testing for Continuous Assessment. Evans Brothers Limited, Nigeria.

Bell, G., H., (1988) Action Inquiry. In J. Hias and S. Groundwater-Smith (Eds.), The Enquiring Teacher –Supporting and Sustaining Teacher Research. (pp. 40 – 53). The Falmer Press, Taylor & Francis Inc., Philadelphia, PA.

Bryce, T., McCall, J., MacGregor, J., Robertson, I., and Weston, R., (1983) Techniques for the Assessment of Practical Skills in Foundation Science – Teacher's Guide. Heinemann Educational Books, London.

Boud, D., Dunn, F., and Hagarty-Hazel, E., (1986) Teaching in Laboratories. SRHE & NFER, Nelson Publishing, Surrey.

Boykoff Baron, J., (1991) Performance Assessment: Blurring the Edges of Assessment, Curriculum, and Instruction. In Science Assessment in the Science of Reform, American Association for the Advancement of Science, Edited by Kulm, G. and Malcom, M., Washington, DC.

Brigance, A., and Hargis, C., (1993) Educational Assessment – Insuring That All Students Succeed in School. Charles C. Thomas Publisher, Illinois, United States.

Carson, T., Conners, B., Smits, H., and Ripley, D. (1989) Creating Possibilities An Action Research Handbook. An unpublished manuscipt, University of Alberta.

Carson, T. and Sumara, D. (1989) Exploring Collaborative Action Research. Proceedings of The Ninth Invitational Conference of The Canadian Association for Curriculum Studies. Jasper, Alberta, May 5 – 7, 1989.

Chalmers, A. (1982) What is this thing called Science? University of Queensland Press, Queensland, Australia.

Chambers, J. and Sprecher, J. (1983) Computer Assisted Instructions: Its Use in the Classroom. Prentice Hall, Englewood Cliffs, New Jersey.

Comfort, K. (1991) Standing Ovation for Performance Testing. In G. Kulm and S. Malcom (Eds.), Science Assessment in the Service of Reform. (pp. 149 – 161). American Association for the Advancement of Science, Washington, DC.

Doll, R. (1982) Curriculum Improvement: Decision Making and Process. Allyn and Bacon, Inc., Boston.

Doran, R., Lawrenz, F., and Helgeson, S. (1994) Research on Assessment In Science. In Handbook of Research on Science Teaching and Learning. Edited by Gaber, D., Macmillan Publishing Co., New York.

Einstein, A. (1954) Ideas and Opinions, Edited by Seelig, C., Crown Publishers Inc., New York.

Eiser, J., R. (1994) Attitudes, Chaos & the Connectionist Mind. Blackwell Publishers, Oxford, UK.

Finch, F. (ed.) (1991) Educational Performance Assessment. The Riverside Publishing Company, Chicago.

Fitz-Gibbon C., and Morris, L. (1978) How to Calculate Statistics. Sage Publications, California.

Fullan, M. (1991) The New Meaning of Educational Change. The Ontario Institute for Studies in Education, Ontario, Canada.

Gagne, R., M., (ed.), (1967) Learning and Individual Differences. Merrill Publishing Co., A Bell and Howell Co.,Columbus, Ohio.

Gipps, C. (1994) Beyond Testing. The Falmer Press, Taylor & Francis Inc., Philadelphia, PA.

Haertel, E. (1993) Educational Assessment: Expanded Expectations and Challenges. In Educational Evaluation and Policy Analysis, 15, no. 1.

Hagarty-Hazel, E., (ed.), (1990) The Student Laboratory and the Science Curriculum. Routledge, 11 New Fetter Lane, London.

Hargraves, N., and Lynch, P. (1988) Experimenting with Assessment: The Practical Examination in HSC Biology Revisited. In Australian Science Teachers' Journal, Vol. 34.

Jeske, G. (1980) Evaluation of Science Practical Activity. An Unpublished Thesis, University of Alberta.

Kahle, J., B., (1979) Teaching Science in the Secondary School. Van Nostrand Co., New York.

Kane, T. (1983) The Oxford Guide To Writing – A Rhetoric and Handbook for College Students. Oxford University Press, Inc., Oxford.

Kemmis, S. and McTaggart, R. (1988) Action Research Planner. (Third Edition), Deakin University Press, Victoria.

Knutton, S. (1994) Assessing Children's Learning in Science. In J. Wellington (Ed.), Secondary Science – Contemporary Issues and Practical Approaches. Rutledge, London and New York.

Kruglak, H., and Wall, C. (1959) Laboratory Performance Tests for General Physics. Western Michigan University Press.

Kulm, G., and Malcom, S. (eds.), (1991) Science Assessment in the Service of Reform. American Association for the Advancement of Science, Washington, DC.

Kutliroff, D. (1975) 101 Classroom Demonstrations and Experiments For Teaching Physics. Parker Publishing Co., Inc., West Nyack, New York.

Layton, D. (ed.), (1990) Innovations in Science and Technology, Volume 3, Published by the United Nations Educational Scientific and Cultural Organization, Paris.

Leedy, P. (1989) Practical Research – Planning and Design (Forth Edition). Macmillan Publishing Co., New York.

Linn, R. (1993) Educational Assessment: Expanded Expectations and Challenges. In Educational Evaluation and Policy Analysis, Vol. 15, No. 1.

Merriam, S., (1988) Case Study Research in Education – A Qualitative Approach. Jossey-Bass Publishers, San Francisco.

Messick, S. (1992) The Interplay of Evidence and Consequences in the Validation of Performance Assessment. Research Report, ETS, July, 1992.

Moore, G. (1983) Developing and Evaluating Educational Research. Little, Brown, and Co., Boston.

New York State Education Department, (1992) Regents High School Examination: Earth Science – Performance Test Manual, Albany, New York.

Osterlind, S. (1991) Scoring Procedures for Performance Tests. In F. Finch (Ed.), Educational Performance Assessment. The Riverside Publishing Company, Chicago.

Priestley, M. (1982) Performance Assessment in Education and Training: Alternative Techniques. Educational Technology Publications, Englewood Cliffs, New Jersey.

Stiggins, R., and Bridgeford, N. (1982) The Role of Performance Assessment in Day to Day Classroom Assessment and Evaluation. A paper presented at the NCME conference, New York, March 1982.

Swanson, D., Norman, G., and Linn, R. (1995) Performance-Based Assessment: Lessons From the Health Professions. In Educational Researcher, Vol. 24, No. 5, (pp. 5 – 11).

Swezey, R. (1981) Individual Performance Assessment: An Approach to Criteria-Referenced Test Development. Reston Publishing Co., Inc., Reston, Virginia.

Swift, D. (1989) Practical Physics for GCSE. Basil Blackwell Ltd., Oxford England.

Tamir, P., (1991) Practical Work in School Science: An Analysis of Current Practice. In B. Woolnough (Ed.), Practical Science. Open University Press, Buckingham.

Tamir, P. (1974) An Inquiry Oriented Laboratory Examination. In Journal of Educational Measurement, Vol. 11, p. 25 – 33.

Weiner, E. and Stewart, B. (1984) Assessing Individuals – Psychological and Educational Tests and Measurements. Little, Brown and Co., Boston.

Welford, G. (1990) Assessment of Practical Work in School Science. In Innovations in Science and Technology Education, Volume 3, Edited by Layton, D., United Nations Educational Scientific and Cultural Organization, Paris.

Wellington, J. (1994) Secondary Science - Contemporary Issues and Practical Approaches. Routledge, London and New York.

Wilberforce, L. (1910) A History of the Cavendish Laboratory 1871 - 1910. Longmans, Green & Co., London.

Zais, R. (1976) Curriculum Principles and Foundations. Harper & Row, Publishers, New York.

A Practicing Teacher's Ten-Minute Guide Into Performance Assessment In Science and Technology Education

By

Paul Wozny

Contents:

Chapter 1 – Performance Assessment Defined

As a teacher of both science and technology education at the high school level, I have heard about this process "performance assessment" quite a bit lately. It seems to be the new rage in assessment these days, yet, how many of us know exactly what is meant by this term? The meaning of performance assessment in the context of this instructional manual is "evaluation of a student or group of students in the process of solving a hands-on task/skill/problem within the direct or indirect view of a moderator/evaluator." One must appreciate that class size will often limit the ability of a teacher to view each individual student during the process of completing a performance assessment task. In this situation of a large class (over 30 students) the teacher may instruct the students to write their process in solving a particular situation in a portfolio/lab book, which may be assessed at a later time.

The definition of performance assessment, which I have provided, is consistent with that of other scholars in the field of education. Stiggins and Bridgeford (1982) share the following definition of performance assessment in the field of science and technology education: "Performance assessment is defined as a systematic attempt to measure a learner's ability to use previously acquired knowledge in solving novel problems or completing specific tasks. In performance assessment, real life or simulated assessment exercises are used to elicit original responses which are directly observed and rated by a qualified judge"

(Stiggins and Bridgeford, 1982, p. 1). Kruglak and Wall, long considered pioneers of the performance assessment process in education, share a similar definition of performance assessment: "In a laboratory performance test the student must deal directly with real apparatus and with an actual physical situation in order to succeed. He must not be able to circumvent the apparatus." (Kruglak and Wall, 1959, p. 12).

Boykoff Barron (1991) shares the perspective on performance assessment that it "becomes the culminating activity of a unit and provides opportunities for students to synthesize their knowledge, make connections and deepen their understanding of major concepts."

I am drawn to the work of Welford (1990, p. 53) in relation to the function of performance assessments in science and technology education: "a worthwhile aim of practical assessment ought to be to ensure that it takes place in the context of purposeful practical work." Welford expands on this premise with the contention that current trends in the assessment of practical work in science are undergoing a transition from "an illustrative, confirmatory or discovery function, with its apparent afterthought of increasing students' interest and motivation, to one which additionally seeks to develop and rehearse investigative skills. Developments in assessment are now offering criteria against which to judge abilities such as handling the variables of the experiment, deploying the strategies of investigation and using the techniques of

measurement, observation, data organization, interpretation and deduction." (Welford, 1990, p. 37).

Performance assessment allows for the teacher a unique opportunity to test the skills of a particular student or group of students to solve a particular problem or demonstrate a skill within the direct or indirect observation of the instructor. You might be asking yourself, "What does he mean by indirect observation?" I personally use both portfolios and video recordings of students completing performance assessment tasks in situations where I am not able to observe each and every student due to the large size of some classes. The video recordings coupled with the student portfolios allow for later assessment by myself or with a group of teachers.

I have used the performance assessment paradigm to check out the abilities of my grade nine science class to construct both series and parallel electric circuits. In this particular performance assessment students came into my classroom during lunch hour, one by one to construct both types of circuits using actual wires, batteries, light bulbs and switches. I was able to observe each student's skill in performing this task and draw some very important conclusions on the effectiveness of my teaching techniques in this domain. Through the hands-on process of problem solving I was able to gain a very clear perspective on the constructs of my students' knowledge and skills in regard to electric circuits after the learning sequence. Yes, as you can well imagine, this

mode of assessment does indeed take time and planning, but if you are hoping to test your students skills in the completion of actual hands-on problems, this assessment paradigm becomes an essential tool in your assessment arsenal.

Chapter 2 – Benefits of Performance Assessment

Why should I use performance assessments in addition to my regular pencil and paper tests, which I currently employ in my science and/or technology education classes? First and foremost it (performance assessment tasks) will enable you the unique opportunity to view your students constructs of the skills and techniques you have attempted to teach them. Moreover, performance assessments allow you to gain insight into the previous skills that the student may or may not have brought to your class.

In the case of an electronics teacher, you might be getting some new students who have recently transferred into your class from an out of province school. These new students might share with you that they were enrolled in an electronics class at their previous high school. As their new teacher, you would like to get some idea on each new student's skill level in the various domains of study in electronics. In this particular scenario, performance assessment becomes your most powerful tool in assessing the hands-on abilities of these new students. This is especially valuable in terms of safety. Will these new students display a degree of safe practice in the electronics laboratory which you

as the instructor are comfortable with? This instructor might choose to set a soldering problem in which each student might demonstrate their skills with a soldering iron coupled with a variety of wire splicing situations. As the instructor directly observes each new student's ability to complete the hands-on tasks, he/she begins to appreciate with a much greater degree of precision, the entry-level skills of each new student. Did the student follow established safety precautions (safety glasses, heat sinks, etc.)? Did the student have knowledge and skills in each of the wire splicing and soldering operations necessary to continue on with the resident students in your classroom, or will they need remedial lessons in order to work in your particular program? Such issues become critical for Career and Technology Studies teachers who are expected to diagnose the abilities of each student entering their particular domain of learning and look at the possibility of advance credit for skills which a student may already possess.

I am particularly appreciative of the benefits which can be afforded by performance assessment tasks in providing a clear picture of the learner constructs. How has the learner interpreted the teacher's instructions and skills, and to what degree can these skills be replicated in a hands-on scenario. The merits of performance assessment in garnering a realistic appreciation of a particular learner's constructs about a particular skill or problem were emphasized in the pioneering work of Kruglak and Wall in their studies of the Cavendish Physics

Program at Cambridge University. The performance assessment tests were first implemented in 1887. The purpose of the performance assessment tasks was in examining a student's practical skills and comparing his/her achievement with that in written tests. These tests lead to the startling realization that students did not necessarily succeed in their performance tests with the same degree of achievement as in their written tests. Moreover, there did not appear to be a strong correlation between written test scores and performance test scores. These results strongly suggested that performance assessments are necessary components of a science educator's assessment paradigm if one is to truly allow each student's abilities to be fairly assessed. An interesting conclusion to the pioneering study in 1887 at the Cavendish Physics Laboratory was that many strong students in the Bachelor of Physics program of studies illustrated "abysmal failures when confronted with a real piece of apparatus" (Kruglak and Wall, 1959, p. 15).

My personal experiences in the utility of performance assessment testing in 17 years of teaching science and technology education has been extremely appreciative of this assessment paradigm. Why you might ask? I am not only able to increase student motivation by hands-on manipulation of laboratory equipment but I am also privy to the examination of the investigative constructs of each student. Much as in the pioneering work of the Cavendish Physics Laboratory at Cambridge University, I have found that many science students truly underachieve

in written test scenarios, yet can be very high achievers in practical hands-on tests. This suggests to me as an educator, that the integration of performance assessment testing must be increased in order to truly appreciate the potentials and abilities of each student. This is particularly important in these days of standardized achievement tests, which may have many teachers teaching to the test and perhaps skipping out on some of the practical skills which may not be covered in the pencil and paper traditional tests. If performance assessment tasks were to be included in the achievement testing process, one could argue that teachers might include more hands-on experimentation in their classes in order to cover the tested material in the province/state wide achievement tests.

I had the great pleasure of speaking one-on-one with Nobel prize winning physicist Dr. Richard Taylor in January, 1998. He shared his experiences as an experimental physicist and stressed the importance of secondary science teachers in providing not only a strong theory base to their instruction, but also the critical ingredient of hands-on experimentation and laboratory apparatus manipulation. Dr. Taylor viewed this process as critical in the hope of creating a more rounded science student who may be more inspired to continue his/her education in science beyond the constraints of the high school experience. Dr. Taylor shared in the fact that his high school chemistry teacher made science not only informative, but also fun coupled with much hands-on

opportunity to bring science to life. Yet, Dr. Taylor was quite critical of his high school physics instructor, whom in his words failed to inspire his students and provided little in the hands-on experience in the physics domain. If it were not for the chemistry teacher who brought science to life for Dr. Richard Taylor, the world may have lost the contributions of this Nobel Laureate who provided so much in the advancement of particle physics. John Dewey would have been proud!

Chapter 3 – Planning a Performance Assessment Test

Before one flies an aircraft to a destination, you must construct a flight plan. In education one must plan his/her students' performance assessment test/task with the same degree of rigor as the pilot of an aircraft, whose own life and passengers' lives depend on his/her preplanning. Why is this preplanning so necessary? First you must assess what it is that should be included as a performance assessment task. Second, you must ascertain if this particular skill/ability will fit within the constraints of the performance assessment paradigm, within the context of your particular classroom. In other words, is this doable! One must appreciate the time variables that will be entailed in this particular performance assessment task in relation to the number of students in your class and the length of time available for the testing sequence. Finally you must set out on ascertaining if a valid and reliable scoring rubric can be created and applied in your particular pedagogical situation.

You are probably saying to yourself "Wow, this performance assessment process is a heck of a lot of work," which of course is absolutely true, but don't fret, the results are worth it. Once you have developed a quality performance assessment task, it will provide a framework for years to come.

Now let's look at the first step in planning a performance assessment test; assessing what it is that should be included as a performance assessment task. This step is critical for it lays out the foundation on the skills that you will examine in a hands-on environment. I would strongly suggest that you refer to the program of studies for your particular subject area and analyze which skills are expected to be delivered in a hands-on format. This will provide a good starting point in choosing which skills you would like to test for in a performance assessment manner. My own experience in providing performance assessments in Science 9, 10 and 20 point towards activities which are expected as laboratory skills in my students. In this regard I am able to find many practical skills which I can draw from to construct the basis for a particular performance assessment. From these practical skills one should identify the specific variables which will be examined in each student completing the particular performance assessment task.

The process of choosing the appropriate procedure is critical in the testing of a particular skill. Will this particular performance

assessment task involve individual or group participation? Does the procedure allow for adequate comprehension by the student of the variables being tested? Will each student have sufficient time to complete the particular performance assessment task? These are critical questions that one should address in the initial planning stages of a performance assessment in order to provide a realistic framework for your particular pedagogical situation.

In my own experiences planning and carrying out performance assessment tasks in both science and technology education, I have found that the key to success in the performance assessment process is in keeping the tasks quite specific to the curriculum/program of studies for that particular subject. The tasks should be coupled with clearly scripted instructional sheets which describe with written text and diagrams exactly what this particular performance assessment entails. Moreover, I have found that providing the scoring system along with examples of what type of performance will result in a particular grade greatly assists the student in appreciating what is expected out of him/her. In many cases these exemplars are given days before the particular performance assessment exam in order to direct the student's focus, study and rehearsal. Naturally, if the performance assessment were testing a novel problem, then the instructor would not likely provide the preceding information.

The time variable is critical to the successful implementation and continuation of performance assessments in your classroom. If students do not have sufficient time to complete the performance assessment activity in one session, they may have the opportunity to share thoughts with other students outside of the class, which would definitely affect the fidelity and integrity of this particular performance assessment task. Students may also become "turned-off" to this assessment paradigm if the first attempt is confusing and troublesome.

The system of scoring a performance assessment task is also referred to as the scoring "rubric." The scoring rubric must be valid and consistent (reliable). Is this particular performance assessment task testing what I think it should be testing and am I able to assess each student's performance in a consistent manner? The consistency issue has long been a major source of trouble in the domain of performance assessment, as can be witnessed in the Olympic games which is often surrounded in controversy with performance assessment events such as figure skating, diving, and free-style skiing. You as the teacher should have a good idea of what grade various levels of performance will result in. I have found that actually providing exemplars of the various performance levels is a good way of keeping my own evaluation of the students' work consistent. In other words, I establish a very specific guideline for the various grades that will be assigned. My experiences as an Industrial Education teacher made this absolutely necessary, as I was

constantly faced with marking students' project work, both in terms of their skills and finished product. In my first years as a teacher I was often troubled by the lack of consistency that plagued the marking of students' finished projects in the shop, which lead to the development of exemplars of finished products at various levels of quality which would not only give the students an idea of what was expected of them, but also allowed myself as their teacher a more consistent mechanism for evaluating the projects. Obviously, this mode of assessment does require considerable effort and time on the part of the instructor, but once the assessment tasks have been developed coupled with the exemplars for the various levels of competency/skill/quality, one is able to use these assessment tasks over and over with greater efficiency.

An example of developing a performance assessment task occurred in my experiences while teaching photography at the junior high school level. One of the activities in the photography section involved the students taking pictures with a 35-mm single reflex camera and adjusting the aperture, focus and exposure times in order to demonstrate depth of field. Each student was to: 1) shoot a set of 10 black and white negatives demonstrating how camera settings could affect depth of field. Each student was provided with a set of exemplars clearly illustrating the levels of skill/quality necessary to a particular grade (in this case from 1 to 10). I then had the students evaluate their own work in relation to the exemplars provided and later evaluated each student's work by myself

along with the particular student. An amazing result was that the students were more critical of their work then I was, yet the repeatability/reliability of the marks was high. This result was in sharp contrast with the consistency and quality of work I was getting from the students before the use of exemplars at various levels of performance.

I then used this approach in my science classes and was astounded at the results from my students. Not only did the students enjoy the hands-on nature of performance assessments, but also they appreciated the opportunity to test their skills in relation to the exemplars that were given to them previously. With my grade nine science classes, I incorporated a set of performance assessment tasks in the construction of series and parallel circuits involving a switch, battery, and two light bulbs. The students were given the chance to view the teacher assembling the circuits at an exemplary level and later tested. Although the students knew in advance exactly what they would be tested on, I found that it allowed me, as the instructor, the opportunity to examine the students constructs of this skill and better appreciate the level of learning which was occurring in my science classes in regard to this particular skill.

I must stress that performance assessment tasks are quite time consuming and you must appreciate that one might have to incorporate group performance assessments in order to accommodate large classes. The group method of performance assessment allows a convenient way of

gathering information on the overall skills of my class in a more time effective manner in comparison to the one-on-one method. The drawback of the group method of performance assessment is that one cannot be absolutely certain on how much each member of the particular group contributed to the completed project. In such group scenarios, I tend to use the results more as an indicator of the overall abilities of the class as opposed to the summative evaluation of each student.

Chapter 4 – Building a Performance Assessment Test

The process of building a performance assessment test/task involves the synthesis of the planning and design aspects of your particular task. For the sake of this condensed guide into performance assessment, I have broken down the process of building a performance assessment task into 7 relatively easy steps:

1) What is it that we want the student to know/do?
2) Under what circumstances will the student's performance be examined?
3) What are the time and cost restrictions?
4) Will this be an individual or group performance assessment scenario and what percentage of the assessment will be hands-on as compared to written?
5) What are the standards that will be expected for the various grades allocated?
6) How will we administer and score the performance assessment task?

7) How will we ensure that my performance assessment tests are reliable (consistent) and valid (testing what it is we think it is testing).

Let's begin with step one - deciding what it is that we want the students to know or do for this particular performance assessment task. This is a critical step, especially in lieu of the validity issue. One must be certain that the performance assessment test will indeed test for the skills/abilities that are in line with the curriculum for that particular course. I use the program of studies as a reference point to choose the skills that I feel are best accommodated by a performance assessment methodology. I then isolate those variables that are most appropriate to the hands-on orientation of the performance assessment paradigm. Both in technical and science education, there exists a multitude of hands-on skills which are essential for the student. Herein lies the value of performance assessments, for it allows the instructor an opportunity to see if her/his teaching methods have related the skills in the manner initially intended. In my electronics course, I use the performance assessment paradigm both for formative and summative evaluation of each student's soldering skills and have had positive feedback from the students during and after this testing procedure. I found that traditional pencil and paper tests simply did not allow the skills of the student to accurately represented in the case of soldering operations.

Step two investigates the circumstances that the student's performance assessment task will be conducted and examined. Will this

occur during regular class time in your usual classroom? Do students have previous knowledge of the exact skills being tested in this particular performance assessment? Will you be the only examiner or will you have other examiners present? Should this activity be done one by one or by groups of students? Could video cameras be utilized for future scrutiny of each student's performance? These questions must be addressed in the building of your performance assessment scenario in order to ascertain its credibility and feasibility. My experience has shown that performance assessment tasks prove most reliable and effective when very specific hands-on skills are being analyzed with a very specific set of scoring rubrics which minimize any uncertainly in the level of achievement by the examinee. The aspect of students viewing other students' work is also an issue in organizing the circumstances of the performance assessment task. One may have to set up work stations which are somewhat separated by some type of physical barrier in order to minimize students perusing at other students' work. As well one could instruct the students that overt viewing of others' work would result in automatic failure.

The use of video cameras has provided the examiner with the opportunity to scrutinize the student's performance at a later date. As well, one is able to garner the skills of other instructors who may cooperate in viewing the performance assessments at their convenience

and share thoughts on the effectiveness and validity of the particular assessment task.

One of the performance assessment tasks in both my science and electronics classes involves the students solving the rate of beam expansion of a helium neon laser beam. I use three laser stations for this assessment and cycle three students at a time. The students write down their hypothesis, theory, procedure, observations, analysis and conclusions. I then grade each student's results at a later date. This system has proved very cost and time effective. I have included this particular performance assessment in the appendix of this manual and you are free to use it. This particular assessment task is suitable for students from grade nine to twelve. It has been a valuable addition to both my science and technology classes.

Step three in the building of a performance assessment task involves analysis of the time and cost restrictions. If time is at a premium, you may chose a performance assessment activity which allows for many students to work simultaneously coupled with each student recording their methods and results which allow for future perusal by the examiner/s. Cost is often overlooked until it is too late. If your performance assessment activity uses costly materials which may not be reused, it simply might not be feasible. Keep it focused and reasonable, both time-wise and cost. I have had limited success with very complex and long winded performance assessment tasks, and now

stay with those assessment tasks which are affordable and reasonable in time allocation.

The bottom line in today's education environment of achievement testing, is that many instructors are minimizing laboratory experiences in order to focus on the theory which is tested in the provincial/state exams. These high stake exams can make or break a teacher's/student's future. This reality forces the instructor to minimize time lost due to the inclusion of performance assessment activities. I suggest that you set up a realistic time-line for the assessment parameters in your particular course for the entire term, and then look at what time is left for the inclusion of performance assessment tasks. You may have to modify your existing assessment outline, but this is essential in appreciating the time variables associated with the performance assessment process.

Step four in building your particular performance assessment activity involves ascertaining whether it will be an individual or group assessment and planning what portions of the assessment activity will be hands-on compared to written. Where a minimum of time and space are available, I find that group assessments which incorporate a major written component are much more feasible than the one-on-one activity which entails a major hands-on component. In an ideal world, one would like the opportunity to test each student individually, but with a class of thirty-five students, this is often not feasible. With a large class, you could have groups of assessment apparatus set at various locations

in the classroom, with students proceeding in a rotating fashion from one station to the other. The students could be told to write down their theories and results, which could then be collected and assessed at a later date. Once again the utility of video cameras is beneficial in this particular circumstance in order to review each individual student's actions.

Budget cuts being faced by many teachers is reducing the amount of physical materials and apparatus in the science and technical education laboratory. It may be necessary to script a problem to your students, or demonstrate it once while the entire class is watching and then have each student record how they might attempt to solve this particular problem. Computer-based simulations could prove to be invaluable in this regard. As in aircraft simulators that allow pilots to train in a much more cost efficient manner, the computer environment could provide a one-on-one economical alternative to the traditional hands-on practical assessment. One could have the components of the performance assessment task manipulated onscreen in a real time environment. This computer-based performance assessment would require considerable programming skills, but may be well within the abilities of certain teachers.

The fifth step in the development of your performance assessment involves the determination of the standards that will be expected for the various grades allocated. This critical step will involve you setting up the

scoring rubric for the particular performance assessment task. I use the following rubric for the majority of my own performance assessment tasks in both science and technical education:

3 Skill demonstrated with complete proficiency.

2 Skill demonstrated with mid-level proficiency with a minimum of instructor intervention.

1 Skill demonstrated with lower-level proficiency with substantial instructor assistance required.

0 Skill demonstrated with no level of proficiency. Constant instructor assistance required.

This particular rubric was adapted from the Northern Ireland Schools Examination Council (NISEC) that has put in considerable research and practice in the delivery of performance assessment tests in their public schools.

As the instructor and designer of the particular performance assessment task, you must also decide how many variables and skills should be judged in the task. In the case of a performance assessment in technical education involving the student soldering a wire splice, one might have the rubric applied to:

1) safety
2) wire cutting and stripping
3) wire splicing skills
4) soldering skills

Of course, your own performance assessment task will have its own set of variables as determined by your particular pedagogical situation. I often find that it is valuable to actually do the performance assessment myself in order to ensure that the assessment is fair and complete. One might also invite colleagues to try out the activity and suggest any improvements.

Step six in building your performance assessment task involves the administration and scoring. Will this test be administered in the regular classroom during regularly scheduled class or will students drop in one-by-one during certain times outside of regular class time? Is the scoring to be done immediately during the activity or will the student fill out a record of their performance with the instructor assessing the record at a later date? Will the performance assessment tasks be monitored by an audio-visual system allowing for future perusal by instructor/s. Is the scoring of the performance assessment task done by one instructor, or will a group of instructors score the task? Do students complete the task individually or in groups? If the tasks are completed in groups, is the mark assigned to the group given to each member of the group, or is there a mechanism to delineate the performance of each student within the group? Each of these questions should be addressed before you attempt the particular performance assessment in order to ensure its feasibility within the context of your particular teaching situation.

The administration of the performance assessment test is a critical aspect that will directly affect its success or failure. Personal experience has demonstrated that practical tests which allow for immediate viewing and scoring provide the best opportunity to appreciate the student's mental constructs and abilities. The downside of this one-on-one testing situation is the time that it takes for larger classes. In the case of larger classes, I will setup a number of workstations so that multiple numbers of students can complete the particular task while I rotate from area to area to view the students. While in the multiple workstation mode, it is necessary for each student to record exactly what steps he/she used to solve the particular problem. This recording process becomes part of the student's academic portfolio that I mark at a later time. The problem with the group approach is based in the inability of the instructor to be all places all the time. In such group situations, I have had greatest success with performance assessment tasks which allow for a high emphasis on the written records by each student as he/she attempts to solve the hands-on problem. Unfortunately, the written record does not provide the examiner with a real time view of the student physically interacting with the experimental apparatus. Moreover, if the student was given information about the exam before its administration, he/she could fake the written report without the instructor having the opportunity to verify the student's actual manipulation of the physical apparatus.

The scoring of performance assessments has been under severe scrutiny due to its potential for unreliable results. If an incomplete or inadequate scoring system (rubric) is utilized both validity and reliability can be compromised. In order to achieve a high degree of consistency in the scoring of a student's performance, one must employ a very specific model of what makes up the various levels of achievement for that particular task. In the case of a grade nine science student constructing series and parallel circuits, this should include the amount of instructor assistance that was required to complete the various stages of the assessment.

At the end of this instructional booklet I have provided a sample performance assessment task which has proved very successful in terms of reliability (test-retest and interrater) and student satisfaction. In this particular task the student is challenged to solve the rate of laser beam expansion for a one milliwatt helium neon laser. The students are provided with a metre-stick, masking tape, and laser target. The scoring rubric includes the domains of:

1) group work,
2) data collection,
3) graphing technique,
4) establishment of relationship,
5) prediction of distance,
6) evaluation of extrapolation,
7) Extra.

In each domain the following scoring rubric is applied:

4 skill demonstrated with full competence with minimal teacher guidance

3 skill demonstrated with high level of competence with occasional teacher guidance

2 skill demonstrated with mid-level competence with occasional teacher guidance

1 skill demonstrated with fair to poor competence with frequent guidance

0 absence of skill demonstrated coupled with the need for continuous teacher guidance

The preceding rubric logic is similar to that shared by the Alberta Education Student Evaluation Branch (1993) which stated that the domains of evaluation in performance assessment tasks in science education should include:

1) initiating and planning
2) collecting and recording,
3) organizing and communicating,
4) analyzing,
5) connecting, synthesizing, and integrating,
6) evaluating the process or outcomes.

In each of these domains the method of evaluation follows a four point scale with level one as the lowest and level four as the highest score. Each level from one to four is explicitly described for each of the evaluation domains. In the domain of analyzing the scoring rubric recommended by Alberta Education follows:

1 correctly identifies patterns within the data; identifies with teacher assistance the various relationships

2 assesses patterns and trends that are conceptually presented by the data; identifies simple cause and effect relationships; identifies with teacher assistance the sources of error in data collection and manipulation; identifies with teacher assistance the effect of errors on results

3 assesses patterns, trends and simple relationships; identifies cause and effect relationships; identifies the sources of error in data collection and manipulation; suggests amendments to procedures and/or data manipulation in order to rectify results

4 assesses patterns, trends and relationships resulting from collected and manipulated data; identifies the sources of error in data collection and manipulation; expresses accuracy qualitatively and/or quantitatively (percent difference), where applicable; identifies the assumptions relating to measurement and/or analysis; determines the reliability of the data.

(Alberta Education – Student Evaluation Branch, p. 9, 1993)

The scoring rubrics that you decide to use should reflect clear differences in the various levels of performance. Have another teacher inspect your rubrics in order to ascertain its clarity and ease of utility. One may also record the students performing the tasks with a video tape recorder and have another teacher evaluate allowing for later comparison in order to gain insight into the reliability of your assessments.

The validity of a performance assessment task is based upon the necessity that this particular assessment indeed tests what it is supposed to be testing. My method for checking the validity of my own performance assessment tasks is based in sharing my assessments with my colleagues and gathering their comments on the face validity of the particular task and comparison with recommended laboratory activities in the relevant program of studies.

The sample performance assessment following this chapter is free for you to use and modify. This particular activity has proven both reliable and satisfying for my students in both physics and technology education. Go forward and be bold in the domain of performance assessment! Your students deserve the chance to show off their skills!

References

Boykoff Baron, J., (1991) Performance Assessment: Blurring the Edges of Assessment, Curriculum, and Instruction from *Science Assessment in the Service of Reform*, American Association for the Advancement of Science, Edited by Kulm, G. and Malcom, M., Washington, DC.

Brigance, A., and Hargis, C., (1993) Educational Assessment – Insuring That All Students Succeed in School. Charles C. Thomas Publisher, Illinois, United States.

Comfort, K. (1991) Standing Ovation for Performance Testing. In G. Kulm and S. Malcom (Eds.), *Science Assessment in the Service of Reform*. American Association for the Advancement of Science, Washington, DC.

Doran, R., Lawrenz, F., and Helgeson, S. (1994) Research on Assessment In Science from *Handbook of Research on Science Teaching Learning*. Edited by Gaber, D., Macmillan Publishing Co., New York.

Kruglak, H., and Wall, C (1959) Laboratory Performance Tests for General Physics. Western Michigan University Press.

Tamir, P., (1991) Practical Work in School Science: An Analysis of Current Practice. In B. Woonough (Ed.), *Practical Science*, Open University Press, Buckingham.

Welford, G. (1990) Assessment of Practical Work in School Science from *Innovations in Science and Technology Education Volume 3*, Edited by Layton, D., United Nations Educational Scientific and Cultural Organization, Paris.

Performance Assessment Task 1
How far is that Laser?
Grade 11 Physics Level

The Challenge:

A Laser beam is directed toward the side of a barn which is a certain distance away from the Laser. The Laser is perpendicular to the wall of the barn. The size of the round spot of light, on the side of the barn is 250 millimeters in diameter. How far away is the Laser from the barn?

Materials:

1 helium neon Laser with a maximum output of 1 milliwatt

paper, pencil, eraser, graph paper, meter stick, metric ruler

Safety:

1) The Laser beam can damage the eye! **Never look** directly **into** the **Laser beam** or stare at any reflections of the beam. Always know where the Laser beam is going and where it ends. Make certain that the beams path will not intercept some person's eyes or skin.

2) If the Laser beam is covering a large distance, keep it close to the ground so that it will not intercept anyone's field of vision. Always keep the Laser beam away from eye level.

3) Use a dull, non-reflective piece of cardboard or other material to stop the Laser beam. You should use a material that will not reflect any light. The best materials are usually of a flat black color.

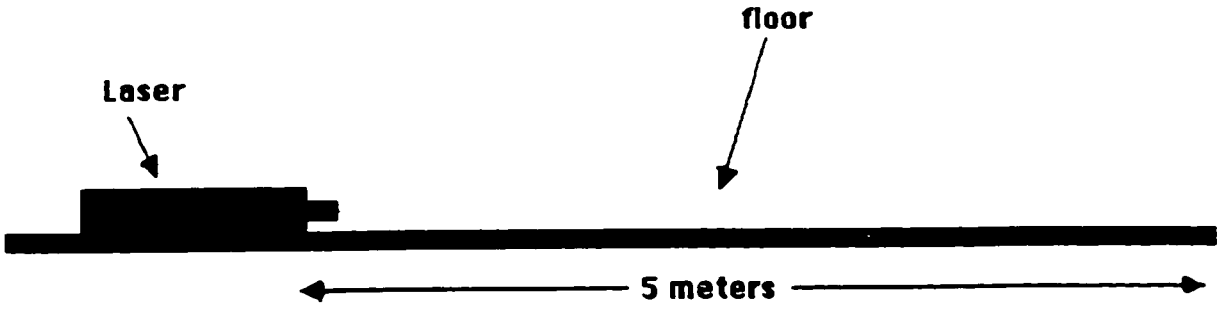
4) A Laser is a fragile piece of breakable equipment. Handle it with care, and do not drop or bump the Laser.

Possible Method of Solution:

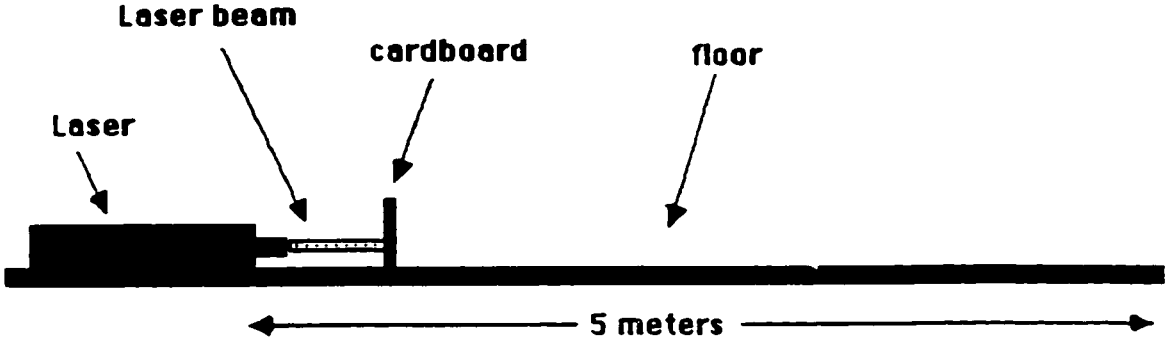
In this experiment you are to find the rate of Laser beam expansion and then extrapolate the distance of the Laser from the barn. You are also expected to graph your data of distance verses beam diameter and calculate the slope of the graph.

Hints:

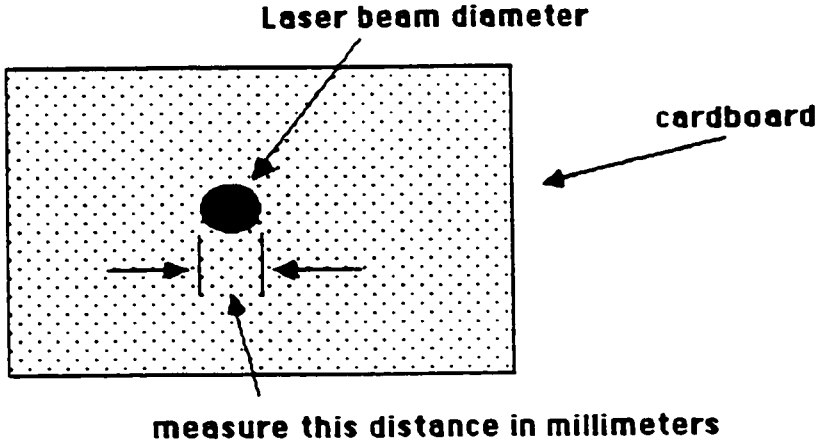
1) Place Laser on the floor, in an area where you have at least 5 meters of unobstructed distance.



2) Take a piece of paper and a pencil, and measure the diameter of the Laser beam at various distances from the Laser. You should take at least 20 measurements.



3) When you take measurements of the diameter of the Laser beam, be certain that you use as much precision as possible. The results you obtain will only be as accurate, as the measurements you take. Make certain the paper is perpendicular relative to the Laser. For bonus marks you could include an estimate of the error in your measurements.



Observations:

Record your data in the following table. Bonus marks will be given for inclusion of possible sources and amounts of error/uncertainty. Include the units of measurement.

Distance	Beam Diameter

Assessing Laser Activity

Assessment Parameter	Scoring
1) Group Work	
2) Data Collection	
3) Graphing Technique	
4) Establishment of Relationship	
5) Prediction of Distance	
6) Evaluation of Extrapolation	
7) Extra	

Total = /28

4 - Skill demonstrated with full competence, with minimal teacher guidance.

3 - Skill demonstrated with high level of competence with occasional teacher guidance.

2 - Skill demonstrated with mid-level competence with occasional teacher guidance.

1 - Skill demonstrated with fair to poor competence with frequent guidance.

0 - Absence of skill demonstrated coupled with need for continuous guidance.

Electronics - Performance Assessments

Task 1 - Draw a schematic diagram of 1 lamp connected in series with 2 lamps in parallel which are all connected to a 6 volt DC source. There must be one switch to turn on/off all three lamps and two more switches to turn on/off either of the two lamps in parallel.

Task 2 - Construct this circuit with the provided materials. Operate the circuit and explain the differences between the series portion and parallel portion of the circuit.

Task 3 - Using the voltmeter of your choice measure the voltage drop across the lamp in the series portion of the circuit.

Task 4 - Assuming that this circuit draws 0.15 amperes of current at 6.0 volts, what would be the amount of energy used by this circuit over 2.0 hours?

Scoring of the Performance Assessment Tasks in Electronics

Points	
4	Skill demonstrated with full competence with no assistance.
3	Skill demonstrated with high level of competence with occasional guidance.
2	Skill demonstrated with mid-level competence with occasional guidance.
1	Skill demonstrated with fair to poor competence with frequent guidance.
0	Absence of skill demonstrated coupled with need for continual guidance.

Performance Assessment Task 2 - Measuring Density

You are to measure the density of the object given to you as accurately as possible. You may use any of the apparatus provided. The following scale will be used in order to determine your score in this activity. Good luck!

4 Skill demonstrated with full competence, with minimal teacher guidance.

3 Skill demonstrated with high level of competence with occasional teacher guidance.

2 Skill demonstrated with mid-level competence with occasional teacher guidance.

1 Skill demonstrated with fair to poor competence with frequent guidance.

0 Absence of skill demonstrated coupled with need for continuous guidance.

Scoring Categories

1) Knowledge of Density Relationship	/ 4						
2) Data Collection Strategy	/ 4						
3) Materials and Apparatus Skills	/ 4						
4) Extra	/ 1						
Total Score	***/ 12***						

Science 10 Midterm Exam Fall 1995
Performance -Based Assessment for Unit 2

Tell students prior to exam they will be doing a lab test on microscopes during their Unit 2/midterm exam.
Make sure you've done the lab to observe onion root tip cells and make sure they know the criteria for a proper scientific diagram.
Run this assessment during Unit 2 exam or midterm exam.
Complexity must vary according to class size and time available.
Set up 2 or 3 microscopes at independent stations so that a few students can do this assessment concurrently.
Leave a few types of slides out .

Demonstrate proper microscope use and **make a scientific drawing** of an **onion root tip cell** by **labelling five cell parts** while on **medium power.** **Show your evaluator** a cell you are going to draw when you have the slide ready under the microscope. **Put your microscope away** when you are finished.
Time limit = **5 min.**

Microscope Use: 5 marks
Mark during class time while exam is being written.
Look for "skill demonstrated or not"
- uncover microscope
- turn on light
- put slide on stage
- on low power, use coarse focus
-adjust diaphragm
- on medium and/or high power, use fine focus
-return to low power
-remove slide
-turn off light
-replace dust cover
Evaluator must look at slide in focus to verify cells observed

Scientific Diagram: 10 marks
Mark after exam is completed.
-title in capitals
-labels lined up to right of sketch in lower case letters
-line sketch in pencil (no shading)
-magnification = objective x ocular = 10x10 = 100x for our microscopes
-name & date in lower right corner
-five correctly identified parts = five marks
eg. cell wall, cell membrane, cytoplasm, nucleus, nuclear membrane/envelope, nucloplasm, nucleolus

Questions/ glitches:
- no talking/ hints
- student may draw cells/ parts not actually visible
-objective marking, not wholistic
-and ??????????????

Performance Assessment Task - Uniform Motion

You are to measure the velocity of an air track vehicle. You will be supplied with an air track, timing apparatus, air vehicle, and starting gate. You are to accelerate the air track vehicle by using the rubber band on the starting gate. You have five minutes maximum to complete this task.

Formula:

Data:

Solution:

Scoring

1 mark for correct formula/relationship
3 marks for appropriate apparatus use
3 marks for correct answer and units

Student	Formula/1	Apparatus/3	Response/3	Total/7

Performance Assessment Task
Is this an acid or base and can I neutralize it?

You are to:
1) use the supplied apparatus and materials to determine whether the unknown solution is an acid or a base;
2) describe to the examiners how you will attempt to do this procedure;
3) neutralize this solution using the materials supplied and demonstrate checks for the neutral property of this solution.

Scoring:

a) description of procedure for determination of acid or base (1 mark)

b) hands-on technique in the determination of the acid or base (2 marks)

c) description and hands-on technique for neutralization of solution (3 marks)

Student	a)	b)	c)	Total/6

Chem 20 Performance Assessement

For your performance assessement you must be able to perform the following tasks:

1. Using an electronic balance as outlined in Appendix C of your book on page 530.

2. Using a pipet as outlined in Appendix C of your book on page 532.

3. Proper technique used to filter a precipitate from the solution in which it was formed as outlined in Appendix C of your book on page 534.

4. Preparing a standard solution from a solid reagent as outlined in your notes from chapter V (Solutions).

You will need to bring your calculator. You will have a maximum of 8 minutes to perform these tasks. All required materials and appropriate instructions will be supplied. You may not ask me any questions !

GOOD LUCK !!!

Chem 20 Lab Exam

Name:_____

I. Calculation of mass. (4 marks)

Formula = (1)
Substitution =(1)
Answer with correct units & sig. dig. = (2)

II. Using the electronic balance.(4 marks)

Place weighing boat on pan & press `ON TARE` button. (1)
Slowly add solid to weighing boat using scupulla until desired weight is registered. (2)
Turn balance `OFF`. (1)

III. Preparing the solution. (8 marks)

Dissolve the weighed solid in approximately 50 mL distilled water. (2)
Transfer to the volumetric flask. Rinsing stirring rod & beaker. (2)
Add enough water to the etched line. Meniscus!(2)
Stopper. (1)
Mix.(1)

IV. Proper use of the pipet. (8 marks)

Rinse the pipet with small amounts of solution. (1)
Place pipet in solution, index finger free. (1)
Squeeze bulb and place on the pipet. (1)
Slowly release bulb & allow solution to rise until volume is above brown line. (1)
If solution enters in bulb. (-1)
Remove the bulb & quickly place index finger on tip of the pipet. (1)
Roll index finger until solution is exactly on brown line. (1)
Exactness of measurement. (1)
Let solution drain naturally into the recieving beaker. (1)

V. Filtering the mixture. (9 marks)

Fold filter paper correctly. (1)
Open & place in the funnel wetting with distilled water. (1)
Tip of funnel is into the beaker & touching side. (1)
Decant most of the solution using stirring rod on lip of beaker. (1)
Keep level of liquid in filter paper at least 1/2 cm from top. (1)
Scrape solid using scraper. (1)
Rinse with distilled water to get all the pp't. (1)
Filtrate is clear. (1)
Remove the filter paper without tearing it. (1)

Chem 20 Lab Exam

Name:_____

Na_2SO_4 --> M = 142.04 g/mol
$Ba(OH)_2$ --> M = 171.35 g/mol
$BaSO_4$ ---> M = 233.39 g/mol

1. Prepare 100.0 mL of a 0.100 mol/L standard solution of sodium sulfate.

2. Pipet 10.00 mL of this standard solution. Show me the pipet before emptying it.

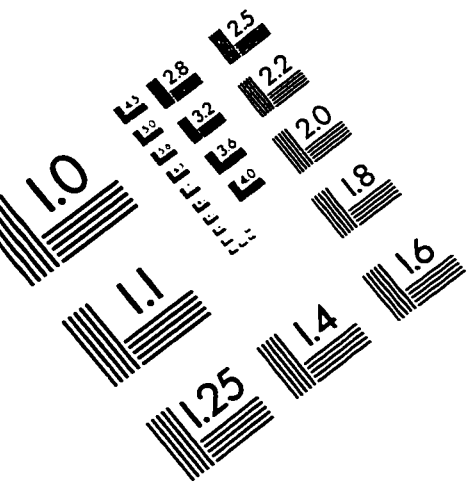
3. Add this 10 mL into the beaker containing the prepared solution of barium hydroxide.

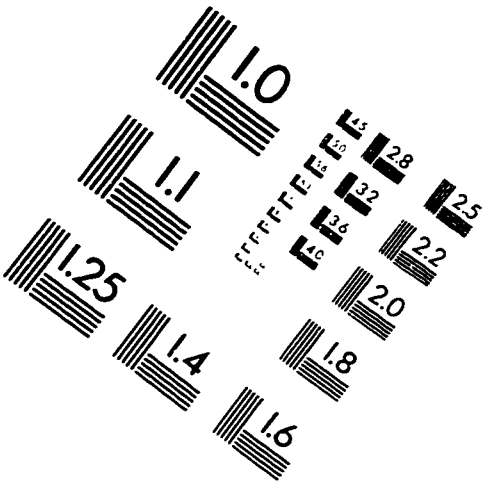
4. Filter the mixture.

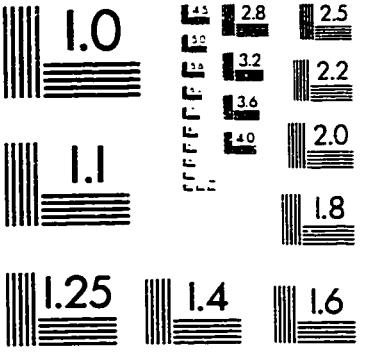
5. Clean the apparatus. Put all waste in the Waste Beaker!

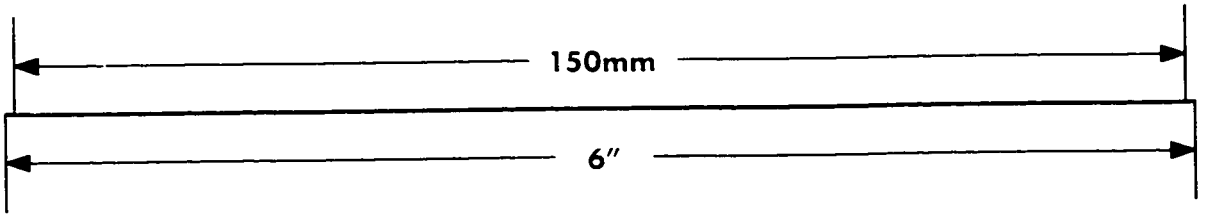
6. Show any calculations in the space below. (Identify what the calculation is for.)

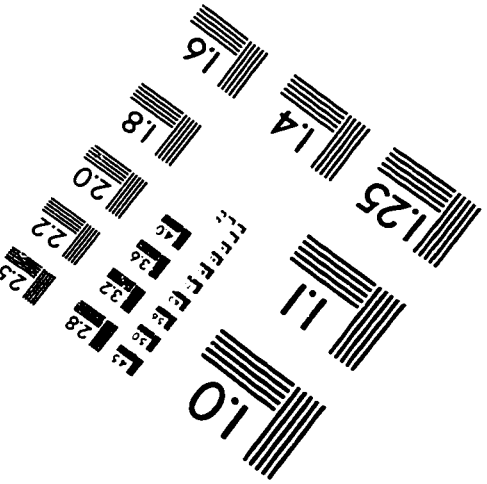
IMAGE EVALUATION
TEST TARGET (QA-3)

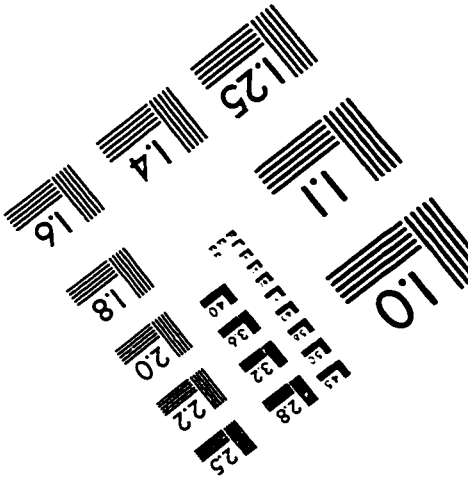












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989