Affecting U.S. education through assessment: new tools to discover student understanding

by

Terry P. Vendlinski
B.S., Political Science, USAF Academy (1979)
M.B.A., St. Mary's University (1986)

SUBMITTED TO THE TECHNOLOGY MANAGEMENT, AND POLICY PROGRAM
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2001

©2001 Terry Vendlinski. All rights reserved

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part.

Signature of Author: [Signature]

Technology Management and Policy Program
February 9, 2001

Certified by: [Signature]

Steven Lerman
Class of 1922 Professor, and Director, Center for Educational Computing Initiatives
Thesis Supervisor

Certified by: [Signature]

Richard de Neufville
Professor of Civil Engineering

Certified by: [Signature]

Judah Schwartz
Emeritus Professor of Engineering Science & Education

Certified by: [Signature]

Ronald Stevens
Professor of Education and Immunology
University of California, Los Angeles

Accepted by: [Signature]

Professor Daniel Hastings
Director, MIT Technology Management and Policy Program
Affecting U.S. education through assessment: new tools to discover student understanding

by

Terry P. Vendlinski

Submitted to the Technology Management and Policy Program
on February 9, 2001 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Technology Management and Policy

ABSTRACT
We may have a serious problem with education in the United States. However, the literature suggests one can arrive at differing conclusions about the efficacy of the American educational system depending on what we measure, how we measure it, when we decide to measure, and why we measure. As will be shown, many testing methods currently in vogue in the United States generate data that can lead policy makers, parents, educators, and even students to arrive at invalid conclusions about schools, teachers, and student ability, especially when evaluating the ability of a student to apply (versus just know) concepts. It was hypothesized that, if applied in a manner aligned with accepted validity standards, modern computer technology could both dramatically improve the accuracy of our inferences, and provide significant new insights into student learning and understanding given present national and California state standards.

As a “proof of concept”, a quasi-experimental, interrupted time series study was conducted using a computerized learning and assessment tool to observe second semester high school chemistry students solving qualitative chemistry problems. The results presented here suggest that:

♦ Without intervention, once a student chooses a strategy to solve a problem, the student will continue to use the same type of strategy (in both the near- and long-term), whether or not that strategy has proven effective. These findings imply we now have the opportunity to both diagnose ineffective strategies as they are developing, and tailor interventions to individual student needs.
♦ Because technology allows us to look both at a student’s answer and how s)he arrived at that answer, we conclude that we can accurately infer whether a student really understands the concepts of a particular knowledge domain. Our findings suggest that, if properly employed, technology can offer new, real-time insights into student understanding.

The paper concludes by discussing the applicability of this research to other knowledge domains, some avenues of future research, and particular pedagogical interventions which the results suggest might be most promising.

Thesis Supervisor: Steven R. Lerman
Title: Class of 1922 Professor, and
Director, MIT Center for Educational Computing Initiatives
Acknowledgments

This dissertation would not have been possible without the help and encouragement of numerous mentors, benefactors and friends. I would like to thank my committee: Professor Steve Lerman whose practical advice greatly helped focus my work and writing; Professor Judah Schwartz whose ability to make learning come alive and desire to improve the quality of education for all I can only hope to emulate; Professor Richard de Neufville for designing a program like Technology Management and Policy without which this research could not have found a home at MIT; and Professor Ron Stevens, a visionary mentor who provided the intellectual, financial and emotional support necessary to undertake the ongoing research begun in this dissertation. Gentlemen, I am grateful for your generous gifts of time and patience.

I would also like to thank those who made the programs, labs, and places I have lived before and during my studies into caring communities: To Pam, Kirky, Helen, and Jud in CECI for whom no need of mine was too small or large to immediately accommodate; to Cail and the TPP staff who constantly went out of their way to support me and the work I came to MIT to do; to Jesse and Rich at Edgerton who daily made an apartment feel like home; to John, Joycelin, Adrian, Pam, and everyone else at IMMEX™ whose intellectual and technical support continues to be invaluable. Jennifer, thanks for your help in my work and your heart in helping me become more accepting of myself and others as the people we are born to be. To Judy, Jim, Anna, Peter, John and Rich at the Paulist Center for your gift of Christian community. To Kathy, John, Tom, Pricilla, Dave and Cindy for encouraging me to begin and continue this work. I thank you for generous friendship over the last decade. To the friends I have grown so close to as a graduate student: Tina, Sebastian, Salal, Ian, and, most especially, to Mike Augustein, Chiz Aoki, and Mark Stoughton. You have encouraged me when I was down, shared my joy when I was up and never let me stop believing this was possible. You all know well that I would not have made it to my defense without your friendship. I am honored to call you friends!

To the Symkowick family – your welcome of me as a brother and dedication to teaching continually remind me of the importance of leaving this world a better place, especially by serving others as an educator. Likewise, Brother Jim – your continual support and encouragement has not only allowed me to achieve an important goal, but has filled the journey with joy. Thank you for going out of your way to be such a dedicated friend.

To my brother Tim, sister-in-law Nancy, and my niece Siena. I am blessed with a family whose unconditional love and support encourages me to follow the dreams I alone might consider impossible.

Finally, I am grateful to God for Life and the two people who allowed that: Life to live in me. To my best friend and father, I thank you for sharing your belief in the importance of hard work, your love, and your faith. Your example, advice, and support have made all I have accomplished possible. To my mom whose short but influential life taught her sons to value and respect all creation. I think of you everyday. It is to my parents that I dedicate this work!

To all of you and many others space does not allow me to mention by name. I am humbly grateful.

Ad majoram Dei gloriam.
Table of Contents

LIST OF FIGURES .................................................................................................................. 7

CHAPTER 1. INTRODUCTION .................................................................................................. 9

CHAPTER 2. THE ADEQUACY AND APPROPRIATENESS OF INTERPRETATIONS FROM ASSESSMENT DATA: WHERE WE ARE AND HOW WE GOT THERE .................................................... 19

2.1 THE DEVELOPMENT OF FORMAL TESTING .................................................................... 19
2.2 THE DEVELOPMENT OF GENERAL TESTING REQUIREMENTS........................................... 21
2.3 THE IMPORTANCE OF CONTENT RELATEDNESS ............................................................ 23
2.4 DETERMINING A "CONSTRUCT" ...................................................................................... 24
2.5 SPECIFYING THE "CRITERION" ...................................................................................... 26
2.6 THE "STANDARDS" ........................................................................................................... 29
   2.6.1 Assuring Content appropriateness ................................................................................. 30
   2.6.2 Assuring Construct appropriateness ............................................................................ 30
   2.6.3 Assuring Criterion appropriateness ............................................................................. 30
   2.6.4 Assuring Analytical appropriateness .......................................................................... 31
2.7 CURRENT (REPRESENTATIVE) ASSESSMENT INSTRUMENTS AND TECHNIQUES .......... 31
   2.7.1 The Third International math and science Study (TIMSS) ............................................. 33
   2.7.2 The National Assessment of Educational Progress (NAEP) ....................................... 36
   2.7.3 The SAT ....................................................................................................................... 39
   2.7.4. The Stanford Achievement Test – 9th Edition (SAT-9) ................................................... 43
   2.7.5 Grades ......................................................................................................................... 46
   2.7.6 Alternative Assessments ............................................................................................. 48
2.8 CONCLUSION .................................................................................................................... 50

CHAPTER 3. ASSESSING STUDENT LEARNING AND UNDERSTANDING WITH TECHNOLOGY .... 55

3.1 THE INTERACTIVE MULTI-MEDIA EXERCISES (IMMEX™) SOFTWARE ....................... 56
   3.1.1 The IMMEX™ authoring module ............................................................................... 57
   3.1.2 The IMMEX™ presentation module ............................................................................ 58
   3.1.3 The IMMEX™ assessment module ............................................................................. 59
   3.1.4 Additional capabilities of IMMEX™ .......................................................................... 62
3.2 STUDY METHODOLOGY ................................................................................................ 62
   3.2.1 Pre-treatment ............................................................................................................. 64
      3.2.1.1 Pre-treatment student survey and results ................................................................. 64
      3.2.1.2 Pre-treatment teacher survey and results ................................................................. 69
      3.2.1.3 Pre-treatment IMMEX™ problem ........................................................................... 70
   3.2.2 Treatment ................................................................................................................... 70
   3.2.3 Post-treatment ............................................................................................................ 73
3.3 CONCLUSIONS ................................................................................................................ 76

CHAPTER 4. IMPROVING INFERENCES OF UNDERSTANDING BY TESTING WHAT WAS TAUGHT ........................................................................................................................................... 81

4.1 ALLOWING SUBJECT MATTER EXPERTS TO EVALUATE CONTENT VALIDITY .................. 84
4.2 ALLOWING THE TEACHER TO FUNCTION AS CONTENT EXPERT ..................................... 84
4.3 THE STUDENT AS EXPERT ............................................................................................. 87
   4.3.1 Item response theory ................................................................................................. 88
   4.3.2 Student performance as content validity measure ...................................................... 91
4.4 CONCLUSIONS ................................................................................................................. 94

CHAPTER 5. DETERMINING WHAT STUDENTS CAN ACTUALLY DO .................................. 97

5.1 THE NEED FOR STANDARDS (CRITERIA) .................................................................... 97
List of Figures

Figure 2-1  Relationship of error types to correlation coefficient and “cut score” in criterion referenced assessment..........................27
Figure 2-2  NAEP assessment content (1999 and 2000)..........................37
Figure 2-3  Relationship of “norm” referenced assessment to minimum competency score (when mean corresponds to minimum).........................................................40
Figure 2-4  Relationship of “norm” referenced assessment to minimum competency score (when mean is lower than minimum).........................................................41
Figure 2-5  Median grade equivalents for sample standardized, “norm” referenced tests over time .......................................................45
Figure 3-1  Problem Space for the Hazmat IMMEX™ problem ..................60
Figure 3-2  Example IMMEX™ “Search Path Map”...................................61
Figure 3-3  Example numerical representation of a student IMMEX™ performance ........................................................................61
Figure 3-4  GPA grade distribution for Ms. Lindsay’s students..................66
Figure 3-5  First semester chemistry grade distribution for Ms. Lindsay’s students.................................................................67
Figure 3-6  Bivariate inter-correlation coefficients (GPA, math and science grades) for Ms. Lindsay’s students ........................................68
Figure 3-7  Box and whisker plot (by period) of median percent correct on Hazmat problems for Ms. Lindsay’s students ..........................71
Figure 3-8  Unit 12 test distribution for Ms. Lindsay’s students ...................72
Figure 3-9  Correlation coefficients between the SAT-9 scores of Ms. Lindsay’s students ...............................................................73
Figure 3-10  Box and whisker plot (by period) of median percent correct on Desperately Seeking Solution problems for Ms. Lindsay’s students ........................................75
Figure 4-1  Concepts covered in Ms. Lindsay’s qualitative chemistry unit .................................................................87
Figure 4-2  Rasch logistics model (problem difficulty = 0) .........................89
Figure 4-3  Rasch logistics model (problem difficulty = 1) .........................90
Figure 4-4  Correlation between student ability by type of Hazmat problem and overall student ability .............................................92
Figure 4-5  Correlation between student ability by type of Hazmat problem .................................................. 93

Figure 5-1  Parameters and range of values tested when building the Artificial Neural Networks (ANN) in this research .......... 108

Figure 5-2  Actual parameter values used for the ANN in this study .......... 110

Figure 5-3  Types of student strategies Ms. Lindsay’s students used to solve Hazmat problems during this study .......... 115

Figure 5-4  Relationship between strategy type a student used to solve his or her first Hazmat problem and the strategy type used most often to solve Hazmat problems .......... 117

Figure 5-5  Relationship between strategy type a student used to solve his or her first Desperately Seeking Solution problem and the strategy type used most often to solve Desperately Seeking Solution problems .......... 118

Figure 5-6  Relationship between strategy type a student used most often to solve Hazmat problems and the strategy type used most often to solve Desperately Seeking Solution problems .......... 119

Figure 6-1  Cross-tabulation of student performance on Hazmat assessment problem when the teacher graded student notes were using a rubric and the researcher graded student’s IMMEX™ performance .......... 129

Figure 6-2  First diagonal of Cronbach correlation triangle (Understanding metrics) .............................................. 130

Figure 6-3  Second square of Cronbach correlation triangle (Understanding vs. content metrics) .................................. 131

Figure 6-4  Final diagonal of Cronbach correlation triangle (Content metrics) .................................................. 131
Chapter 1. Introduction

Americans, throughout our history, have seemed preoccupied with rank. We want to be first in everything. As a nation, for example, we take great pride in being the first and only nation to have ever put men on the moon. We demand the finest health care available and the highest standard of living. We demand the "strongest" military, and require the latest in technology. Our television, radio and print media are peppered with statements like, "When only the best will do" or "Quality is job one." We have created a culture and society that reinforces this "keep up with the Joneses" mentality.

As the new millennium dawned, the United States enjoyed one of the most prosperous and prestigious periods in its history. One of its chief rivals on the world stage, the Soviet Union, had crumbled and the U.S. economy was in the ninth year of strong economic growth. In spite of this prosperity, there were still pressing problems. One of the most often mentioned problems in the 2000 US Presidential Campaign was the perceived poor performance of America's schools. As in most other things, Americans wanted a first class educational system, and the polls of the time suggested (by and large) Americans felt their schools were not delivering the goods. (NCTPP 1990; Berliner and Biddle 1996) But this was not a new problem. The belief that American public education had failed the citizenry seemed a common refrain in the last half of the 20th century, especially when some unexpected event suggested America's students were "less than the best."

When the USSR launched Sputnik in 1957, the people of the United States were fearful. Not only did they feel they had "fallen behind" in what was later to be called the "space race," but this technological leap would almost assuredly afford the Soviets an opportunity to hold America hostage to their growing arsenal of atomic weapons. How had such a thing happened? Much of the blame for this technological impotence was attributed to the lack of math and science excellence in US primary and secondary education (Anagnostopoulos and Williams 1998). The states and nation undertook large-scale efforts to reform the curriculum, especially the math and science curriculum. Yet, by the end of the century the many "reforms" that followed seemed to have produced meager results at best.

Barton notes that, "Any discussion in the United States of how well American students are educated and how to 'reform' education comes -- either quickly or eventually -- to testing." (Barton 1999) If one is to believe them, data from 25 years of the National Assessment of Educational Progress (NAEP) shows average scores in both math and science have remained fairly constant (NCES 1996). Scores from the Scholastic "Aptitude" Test (SAT)¹ and the American College Test (ACT), however, showed substantial decreases during the same period (Hashway 1998). These continued poor performances by American students prompted then-President Bush to meet with the nation's governors in 1990 to discuss the future of American education. Together they established national educational goals, amongst which was that the US would be "first in the world in math and science by the year 2000." (USDoEd 1993) As with previous

¹ The Educational Testing Service no longer refers to this test as the "Scholastic Aptitude Test," rather it is now known only as the "SAT."
"reforms" little seemed to come of this pronouncement by the new millenium. In fact, the century’s last international study of pre-collegiate math and science ability was another major disappointment for many in the U.S.

In part this test, the Third International math and science Study (TIMMS), revealed that American high school students were among the industrial world’s least able students. (IAEA 1994) Although similar tests had produced comparable results in previous years, these results had been criticized and explained away because of perceived differences in cultural and educational priorities between the US and other countries. The TIMMS supposedly corrected or statistically controlled for those differences. The results prompted the statistician who coordinated the American portion of the test, William Schmidt, to claim that the TIMMS “burst another myth” about the adequacy of American math and science education. (Bronner 1998) In the wake of this test came renewed calls to improve the American education system or face the consequences of a lower standard of living and quality of life as the better educated students in other nations enter the workforce. (USDoEd 1983; Shaw and et. al. 1997)

In light of decades of supposed failures in our educational system, it seems paradoxical that the U.S. economy and industry have remained successful in some of the most educationally demanding fields of human endeavor. While true that American high technology firms have occasionally sought increased immigration quotas in order to accommodate their need for more skilled labor, the country continued to send spacecraft out of the solar system, develop Very Large Scale Integrated (VSLI) electronics, and add to its list of Nobel Laureates. In addition, the number of its citizens with post-secondary degrees continued to increase steadily. (_____ 1998) Although numerous antecedents might explain some of these outcomes, it seems certain that such widespread success would not be possible if the American educational system was not imparting important basic facts and skills to its students. (Berliner and Biddle 1996) Nevertheless, aside from the continued successes of American industry and the fact that increasing numbers of the population obtain advanced degrees, America is unable to gauge the effectiveness of its educational system with any consensus of certainty.

It is ironic that while decades of student test results exist, these data provide little information on the extent of student understanding of or how well students can apply what they know. We have difficulty measuring the effectiveness of schools because we have yet to decided on what constitutes effectiveness and present tests, for the most part, measure what is easy to measure and what is easy to quantify. (Wiggins 1993)

The history of American educational reform in the later half of the twentieth century appears to be a series of knee jerk reactions to the latest pedagogical trend or piece of bad news. Because of their transitory nature, some would call reforms such as “whole language”, “new math”, and “back to basics”, to name but a few, “fads.” As early as the 1950’s however, educational researchers advocated the development of the means to authentically evaluate these “fads.” (Bloom 1956) In fact, Bloom lamented that with all the data at its disposal, it was unconscionable that the U.S. education system had not yet developed the means to do just that.
The difficulty, however, does not seem to be that American psychometrists have difficulty gathering or analyzing quantitative data, but that as a society, we have yet to decide why we are measuring what we measure. In the years from 1975 – 1985, the number of state testing programs grew from one to 34. By 1989, every state had some sort of mandated test program. Despite this intensive testing program however, The National Council of Testing and Public Policy (NCTPP), finds these tests have done little to affect the curriculum of schools or the abilities of students, and have little correlation with future job performance or college success. In fact, the NCTPP reports that dissatisfaction with public schools has increased over the same period of time. (NCTPP 1990) Linn reports that while other state standardized test scores increased during this decade, scores on the National Assessment of Educational Progress (NEAP) were more modest, which raised concerns that test gains were not generalizable. (Linn 2000) American policy makers must first decide why they are measuring in order to know what to measure and how to evaluate it. For example, Goals 2000 specified a succinct goal of “first in the world,” but is this really the desired outcome? What does “first in the world” mean? If every other country decided to graduate secondary students because they could sign their name, would Americans be happy that their graduates could sign their name and write the date? After all, American graduates would then be “better” than the graduates of other nations. The point of the obviously silly metaphor is that merely comparing one individual or a group of individuals to another individual or group of individuals may tell you little about the ability of either group.

Such normative comparisons seem to beg for misuse. In his now famous Lake Wobegon Report (named for the fictitious town in Wisconsin where “all the children are above average”), West Virginia pediatrician John Cannell showed that the vast majority of school districts and all the states were scoring “above average” on nationally normalized, standardized tests. Aside from telling us that the test might need re-normalizing, what inferences can be made from the data produced by such tests? In fact, it seems clear that the President and Governors really wanted something more than just a stellar ranking. In addition to being “first in the world” they wanted students that could “demonstrate proficiency in …science; and …use their minds well.” (USDoEd 1993) This objective gets us closer to what the nation really seems to desire, although the terms “demonstrate proficiency” and “use their minds well” are still somewhat nebulous. The Goals 2000: Educate America Act further specifies the need for “students who demonstrate the ability to reason, solve problems, apply knowledge…” (Sec. 102 (3)(b)(ii)). Ted Sizer also clarifies the goal when he observes that what policy makers and stakeholders want from the American educational system are not high school graduates who have passed a certain number of “Carnegie units.” Nor do we want those who have merely earned sufficiently high tests scores. Rather we want graduates that can “exhibit” their accomplishments at a satisfactory level of expertise. (Sizer 1984) And Gardner forcefully argues that in a time when no one can possibly learn even a small amount of extant knowledge we need to acknowledge that people learn and express that learning differently. (Gardner 1997)

Along with the realization that changes are required in the way teachers teach and students learn, we must also acknowledge that the way we test students must also change.
The current multiple choice test is over relied on, lacks public accountability, can lead to unfairness in the allocation of opportunities, and too often undermines social policies according to the NCTPP. (NCTPP 1990) In order to achieve true reform, tests must become assessments. Wiggins defines assessment as “a comprehensive, multifaceted analysis of performance.” (Wiggins 1993)

Hashway recounts how, even these sorts of attempts at reform have, in the past, fallen short of what is really desired. He recounts that early in the century, an American student was rated on how many correct responses he could give in a fixed period of time. The proportion correct was assumed to reflect how well the student would have done in a longer period of time, on all the questions about a domain. The popularity of Taylor’s “time and motion studies” introduced a more quantitative (or “scientific”) quality to testing (Hashway 1998), and the marketing of Binet’s scale as a measure of general intelligence by Goddard, Terman and Yerkes, went far to convince the average American that intelligence could be encapsulated in a single numerical score. (Gould 1996) For over 70 years this quantitative dimension has become the sine qua non of American student evaluation. In the early sixties, Glaser (Glaser 1963) and others called for reporting student scores that were measured against some performance. Bloom (Bloom 1984) and others have interpreted this as the need to establish a cutoff score and teach students to exceed that score, that is to develop “mastery” in a particular academic area. Nevertheless, even this has produced a flurry of claims and counterclaims about the effectiveness of these mastery learning programs (See for example, (Aaderson and Burns 1987, Guskey 1987; Slavin 1987; Slavin 1987)). Still others have tried to make test scores meaningful in relationship to some specific content area. Novick and Lewis (Novick and Lewis 1974), for example, propose a continuum that represents the degree of success a student would experience if confronted by a representative event from that content domain. It is this last proposal that seems nearest to what we really want to know about our students but is, at present, still largely underdeveloped beyond specific qualitative methods.

In this dissertation, I make the assumption that the American public and America’s policy makers really want to know who understands the “things” schools, employers, and other citizens consider “important.” These “important things” might be the basic tools of good citizenship, the knowledge necessary to optimize the choices of a high school graduate, or some other set of cognitive abilities. I will demonstrate how such an analysis might proceed using one very small part of the “standard” American High School chemistry curriculum as an example. As such this research represents a “proof of concept” and is intended to demonstrate how educational assessment, in general, might be improved by the sensible application of computer technology. I will argue that, heretofore, we have lacked the appropriate testing methods necessary to reach substantiated conclusions about what students know, and how we can help them understand and apply concepts to real-world problems. I will focus on how the society might improve the accuracy of the inferences it makes about an individual’s understanding of some of the concepts and ideas he is presently required to learn in a specific setting and demonstrate the improvement we can expect from bettering this accuracy. While I use the Interactive Multi-media Exercises or IMMEX™ software, my findings suggest any software that
offers similar capabilities (e.g. it is adaptable by the teacher, it is designed for student-centered problem solving, it records student work, etc.) should produce results similar to those described here. Unfortunately, I have been unable to find any other software with the range of capabilities of IMMEXTM.

America has once again changed the demands it places on the data that results from testing. In the early 1900's we demanded proof that a student had mastered certain basic cognitive or manual skills. Beginning in the 1920's the focus decidedly favored comparison of students to the "average" (or norm) student. Since the early 1960's and with gaining force in the 1980's and 1990's there have been increased calls for assessments designed to measure a specific set of criteria in order to produce the data to evaluate student understanding. Throughout the century, though, teaching and testing have changed relatively little prompting Papert to observe that if a teacher from the 1800's would walk into a classroom today, they would find little had changed since they practiced their art. (Papert 1993)

Perkins discusses understanding as the ability to build rich cognitive architectures, dig out knowledge, offer explanations and actually use knowledge to solve problems. (Perkins, Crismond et al. 1995) Perhaps this definition offers a richer insight into what Goals 2000 called "use their minds well." The general thrust seems clear; students must not only demonstrate they have knowledge but that they can also apply this knowledge in authentic situations. By authentic situations I mean situations one could conceivably face or be required to resolve at some time in their life. As will be demonstrated, the literature is resplendent with criticisms of current tests that lack any relationship to how a student might actually apply the knowledge they learn in school. This is no true of how the states test in other areas. Not one of the fifty states, for example, assesses a new driver’s ability to drive with a written test. The written test evaluates memorized knowledge of rules and law. The actual assessment of one’s ability to drive takes place behind the wheel. Maker cites no less than five studies indicating that our current knowledge tests provide little useful information about how well students will do in careers or high-level, non-academic settings. (Maker 1994). Crouse and Trusheim detail the futility of trying to predict college success from a single written test like the SAT. (Crouse and Trusheim 1988) As was mentioned, America has decades of written test results. What we don’t have are results that allow us to infer how well students can actually apply what they know (in a new course, a new job, a new school, or in other actual life situations). “You get what you assess...[and] you don’t get what you do not assess.” (Resnick and Resnick 1990) In order to know if a student understands (as Perkins describes it) or can “use their minds well”, we need data that will allow policy makers, teachers, parents and the student themselves to make valid inferences in two different areas. We must have the ability to assess both how much content a student has learned and how much of that content they can apply proficiently to resolve a problem or achieve a goal. (Wolf, Bixby et al. 1991) Maker terms evaluating only the first aspect rather than both aspects of this duality the difference between testing and assessments (Maker 1994), and this is precisely the tack Bloom’s Taxonomy of educational objectives took almost 50 years ago. “Knowledge in and of itself would seldom be viewed as the final goal of education” states Bloom. “It is generally expected students will be able to apply what they learn to new problems.”
(Bloom 1956) Knowledge is, in fact, only the first of six stages of learning for Bloom. But he is quick to point out that,

"Many educators value knowledge because it can be taught (or even learned) relatively easily, and we often believe it is learned simply because it is presented. Because of this it is frequently emphasized as an educational objective out of all proportion to its usefulness or its relevance for the development of the individual. Requiring a student learn certain knowledge assumes that the student is likely to be able to use this knowledge in the future." (Bloom 1956)

But information is growing too fast. We are unable to anticipate the knowledge a student will need, because that knowledge is not yet known, and even what is currently known is too vast to be learned in its entirety. (Gardner 1997) Although test practice has remained unchanged from an era when it was considered enough for schools to teach mastery of routine skills, new research suggests it must change. In 1922 Thorndike advocated the "exercising of the bonds between stimuli and response...and stamping out inappropriate responses." (Thorndike 1922) This is termed decomposability and along with decontextualization (i.e. that a skill is a fixed entity that will take the same form in any setting), these concepts have formed the basis of educational testing for decades. However, Resnick cites current research that claims both decomposability and decontextualization are invalid. (Resnick and Resnick 1990)

That the structure of the curriculum bends to serve assessment should also be understood. As more states impose high stakes tests that rank students on their ability to recall isolated fragments of content with little or no feedback on strengths or weaknesses, the state not only fails in its responsibility to teach but it also drives curriculum away from teaching for understanding and toward memorizing for a test of recall. As an analogy, Madaus cites the effect tomato picking equipment had on the taste of tomatoes. Tasty, thin skinned tomatoes, were bred out of the growing lines in favor of the thick skinned, but less tasty tomatoes that could be harvested by machine. Testing what is easy to test does not necessarily produce the data we need to make informed inferences. Some believe that Americans are gradually becoming more aware of the negative results of using tests (not assessments) as instruments of policy. (NCTPP 1990) With this awareness comes the realization that current modes of testing cannot provide the information necessary to make inferences about what a student understands.

As more and more time in the curriculum is devoted to “teaching to” high stakes tests and less to developing understanding, demands for new modes of assessment are likely to increase. When that occurs, the central question will be “then how can we assess understanding?” It should be noted, that it is not my intention to argue that we abandon all multiple choice or norm-referenced tests. Valid tests have proven costly to develop in terms of time and money. Publishers of high stakes tests are well connected politically and have invested large sums of money to convince both policy makers and the public that their product is necessary. (Crouse and Trusheim 1988) For these reasons alone pencil and paper, high stakes multiple choice exams are unlikely to pass away anytime soon. These tests are also easy for time strapped teachers to administer and evaluate. More appropriately, Wiggins and others suggest that multiple test formats must be used
to validly infer student understanding. (Wiggins 1993) This would argue for including multiple choice as part of a mix of assessment tools. But using only tests that evaluate a student's ability to recall knowledge seems likely to produce invalid inferences about the very thing society seems to be most interested in – how well a student understands a concept. Gardner's claims of the multiple intelligences (Gardner 1983), Cole's (Cole 1996) and Rogoff’s (Rogoff 1998) argument that context must be taken into account, and Norman’s suggestion that we need to have multiple representations of a person's mental model to properly represent what they are thinking (Norman 1993) not only argue for numerous test and assessment forms, but for new assessment forms. Wiggins and others provide exemplars of what these new assessments might look like. In general these assessments allow students to compare their own performance to an exemplary performance or criterion and to objectively see how they measure up. The feedback is immediate. In addition, the performance usually involves the production of an authentic piece of work (either a process or product) rather than a listing of disjointed facts in a predetermined order. (Wiggins 1993)

Many would argue that the primary reason testing hasn’t changed in the last 100 years is because the pencil and paper, multiple choice tests currently in use are cheap to administer and cheap to grade in terms of both time and money. They become even less expensive when their costs are amortized over hundreds of thousands of students. In addition, one can point to the reliability and validity studies made on these tests that suggest they allow us to make statistically sound inferences about what students know. Nevertheless, the requirements on current testing instruments have changed. As these tests have been pressed to provide different types of information, we have stretched them well beyond their original designs. More importantly, it will be argued that the type of data they provide can encourage us to make invalid inferences about how well students understand concepts or can use the information they possess.

Inaccurate interpretations of, or the lack of data over, the last half of the twentieth century have led many to argue about a crisis in American education, and to propose "fixes" to the system. There is concern that we are “falling behind” or that students won’t be able to compete in the world of work they will face in the future. The validity of these conclusions is only as sound as the data upon which such conclusions rest. If the data is limited or flawed, both our conclusions and the interventions we make to "correct" the shortfalls in such a system might also be inappropriate. In arguing for the creation of a national assessment, Tyler observed this very thing. In 1966 he stated, “To make these decisions, dependable information about the progress of education is essential... Yet we do not have the necessary comprehensive dependable data; instead, personal views, distorted reports, and journalistic impressions are the sources of public opinion.” (Tyler 1966) As Linn, et al. report, when it came time to develop such a test, the test developers had some trepidation. This was especially true after less than stellar reviews of the initial versions of the pencil and paper test. (Linn, Koretz et al. 1996)

Many believe the advent of powerful computer technology offers us the opportunity to work around some of the limitations of pencil and paper tests and, by virtue of the large
storage capacities and rapid data processing abilities of these machines, to gain new insights into student learning and understanding.

The research question addressed in this dissertation is this: Will the use of modern technology to create assessments more in line with present standards, current assessment needs, and the validity frameworks developed over the last half century, offer more accurate insights into student learning and understanding than presently possible?

I hypothesize that such computerized, authentic assessments will:

- **H1**: Improve the accuracy of our inferences; and will
- **H2**: Provide significant new insights into student learning and understanding.

Given the effect assessment has on curriculum, the policy implications of the answer to this research question are profound. When teachers are better informed of the learning progress and difficulties of their students, they can make better decisions about what a student needs to learn next and how to teach that material in a manner that will maximize the student's learning. (Fuchs 1995) Assessment drives the American school curriculum, (see for example, (Resnick 1987; Shepard 1993; Wolf 1993; Wiggins 1996)) so we need to ensure that assessment is both providing the data we need to make valid inferences about what our students know and are able to do, and that it is driving the curriculum to teach what is really important. As the NCTPP makes so clear, new assessments must promote the talents of all people by ensuring that the inferences made about individuals or institutions on the basis of test performance are accurate. This goes to the heart of test validity and the fairness of how decision makers allocate the limited resources at their disposal. Ultimately, it determines how we resolve the pressing educational “problems” mentioned in the opening paragraph of this chapter. Bobrow argues that we must interpret the problem, specify the goals, identify the information needed, gather the information, develop alternatives, and assess / compare alternatives to develop sound policy. (Bobrow and Dryzek 1987)

In the pages that follow, I will review the major contributors to our current thinking about assessment in the field of education, and I will consider the validity of assessment methods currently popular in the United States. Next, I will introduce an exemplary sample of computer software as a representative of a cutting-edge computerized learning and assessment tool. Within this framework, I will further develop my hypotheses of how this particular tool might improve our inferences about the understanding of students, the methodology I employed to test such hypotheses, and the data resulting from employing that methodology in an American high school classroom. In the three chapters that follow, I will build the case that this technology (and by implication similar technology employed in a similar manner) allows inferences about student learning which are more accurate given present standards. Specifically, I will show how the technology allows assessments that are more closely aligned with what student have learned and how they have learned it (content validity), that are better predictors of future performance (predictive and concurrent criterion validity), and that more accurately measure student understanding (construct validity). More importantly, I will demonstrate how these
improvements offer important new insights into student understanding and suggest interventions that might dramatically improve student learning. In the final chapter, I look at what I believe to be, the important technological and policy considerations that allow for more valid inferences about student understanding. I conclude with a discussion of where the research can proceed from here, and some of the pitfalls to avoid in applying assessment technology in the classroom.
Chapter 2. The adequacy and appropriateness of interpretations from assessment data: where we are and how we got there

Modern computer technology seems to offer us the possibility of improving our inferences about student learning and understanding. Specifically, the tools that are made possible by this new technology could give us new insights into a student's thinking, knowing, and understanding. Might these improved insights allow us to propose ways to help individual students understand concepts more deeply and apply this understanding more effectively to real world problems or challenges? Essentially the answer to this question and the way we reify the possibilities grow out of the validity of the inferences we make from the data generated by test measurements. Therefore, my research question, analytical methods and results are couched in terms of test validity. As will be seen, our failure to attend to this attribute of testing has constrained our curriculum and the way we deliver it. Moreover, because of our propensity to reach conclusions based on data never meant to address such outcomes, student learning and public education policy can be adversely affected. By using the structure of validity to assess the use of computer technology, I hope to accomplish two things: First, to identify the bounds of current testing tools and, therefore, to suggest where new tools are needed. Second, to suggest how new technological tools like the Interactive Multi-media Exercises software might be employed to help fill the gap between what we have and what we need to help educators, parents, students and policymakers make the decisions they need to make in order to improve student understanding across the curriculum.

The implications of this research, therefore, go far beyond a demonstration of the validity of the inferences from a data set generated by a specific computer technology. Indeed, the inferences we can make from observing how a student solves a problem or resolves a dilemma suggest we can move beyond the walls imposed by many of our current testing methods. My results suggest this tool could change the nature of US education. By giving teachers a tool that can suggest curricular changes (i.e. formative tools), we improve their ability to diagnose and treat what is lacking in student achievement. We are able to truly assess the understanding of students—both content mastery and their ability to apply that content in authentic situations. Finally, we can change the nature of the larger educational system itself from an uncaring and punitive evaluator to a demanding, but helpful coach. As a "coach" the system can not only accurately assess an individual's ability, but it can also provide the feedback and regimen necessary to move each of its charges toward higher levels of learning. First I review the development of thinking about testing and the inferences made from current tests. Then I will discuss some of the problems with the inferences often made from current widespread practices. Finally, I address what the literature suggests as an additional source of data from which to make inferences.

2.1 The development of formal testing

Formal tests can be traced at least as far back as the emperors of China in about 2200 BC. Originally these tests were tests of performance to choose government officials. Unlike the systems prevalent in most other nations of the time, Chinese bureaucrats were chosen, promoted, or dismissed based on performance, not their position in an aristocracy. As time passed, the assessments became more embellished and eventually included
geography, law, military subjects, mathematics, writing, agriculture and music (among others)! The tests included questions that involved both proficiency and knowledge. Hashway and others suggests that testing changed little until the middle 1840’s when Horace Mann argued that test scores, which were usually determined from recitation and long essays given by individual instructors, needed reliability and objectivity in order to be useful. (Hashway 1998; Heubert and Hauser 1999) In fact, the cousin of Charles Darwin, Sir Francis Galton proposed measuring intelligence by a person’s sensitivity to a whistle or their ability to accurately predict weight. In America, Cattell used ability to feel pain and the amount of pressure with which a person could squeeze an instrument as predictors of intelligence. Although Sternberg, et al. cite Wissler (1901) as finding no relationship between any of these “intelligence measures” and college grades (Sternberg, Forsythe et al. 2000), the statistical and psychological techniques necessary to accomplish Mann’s vision were not to become widely known and available until the early 1920’s. These techniques were rooted in the work of educators like Thorndike (Thorndike 1922), and the adaptation of Alfred Binet’s intelligence tests to find officer candidates for the Army by Terman and Yerkes in World War I. (Gould 1996)

In fulfilling his mandate from the French Minister of Public Education to “identify those children whose lack of success in normal classrooms suggested the need for some form of special education” (my emphasis), Binet developed what was probably the first formal educational testing program in modern times. (Gould 1996) In a decision that was to have a profound effect on testing for the next 100 years, Binet decided to test everyday reasoning and life skills (i.e. some “innate” learning ability), but not “learned skills” that could change depending on teacher or classroom. He did this to avoid classifying as needing special attention a child that had just been poorly taught. Because of his mandate, Binet also decided to compare individuals to the “normal” student. It should be noted here that, unlike many of his successors, Binet consistently expressed the belief that intelligence was far too complex a construct to express as a single number. (Gould 1996) Nevertheless, this test and the idea of referencing individuals to some normative measure of achievement was far too compelling for many to avoid. The Army’s Alpha and Beta tests epitomize this “blind empirical” approach to testing. That the results were appropriate because they came from a test that measured what it said it measured was a commonly held belief of the day. As an example Shepard cites Guilford as stating that, “a test is valid [to measure] anything with which it correlates.” (Shepard 1993)

There are obvious problems with defining the appropriateness of testing in this way. First and foremost is the statistical fact that, by random chance, the results of testing are bound to significantly correlate with something else occasionally. In addition, this type of definition does not account for intervening variables. In effect, these intervening variable(s) explain some or all of the variability in the outcome, but their effect on changes in the outcome given certain conditions remains unacknowledged and probably unknown. Another major problem with the “blind empirical” approach is that although two variables correlate, without some logical underlying theory as to why such a correlation exists, there is no hint as to the “direction” of correlation or the reason why the variables correlate. In interpreting the World War I Army Alpha and Beta results, for example, Yerkes explained away the correlation between intelligence and the amount of
schooling ("men with more innate intelligence spend more time in school"). Similarly, he dismissed the correlation between health and intelligence ("low native ability may induce such conditions of living as to result in hookworm infection."). (Gould 1996) Recall that these tests were only supposed to measure "native intelligence" and Yerkes was determined to preserve that interpretation. Looking back on these rationalizations one easily concludes that more plausible explanations for the data existed. The point is that statistics alone cannot justify the interpretation of a correlation. The models must also be founded on explicit theory. Even more important is the warning made to all beginning students of statistics. The mere correlation between two variables does not provide sufficient grounds to assume causation. In fact, any variable that has been increasing over the last few years will correlate to some degree with any other variable that has been similarly increasing over the same time period. For example, the number of people who ate tomatoes in the year 2000 and the number of people who died that year both increase as time passes from January 1st to December 31st. We cannot, therefore, conclude that eating tomatoes causes death. We should probably not even assume that eating tomatoes effects the death rate. However, the correlation coefficient is easy to calculate and that, in part, explains its appeal and wide-spread use. Nevertheless, caution must be exercised. As Gould notes, "assuming that the mere existence of a factor, in itself, provides a license for causal speculation" is something others have warned us against. "But our Platonic urges," Gould concludes, "to discover underlying essences continue to prevail over proper caution." (Gould 1996) The idea that the mere correlation between two variables reveals profound information on which to base inferences dies hard. In the late 1980's, Anne Anastasi still finds it necessary to reject the blind empiricism prevalent in the field (Anastasi 1990), and schools of education still teach that a test is valid if it "measures what it says it measures."

While the excesses of the Army Alpha and Beta testing must be viewed in the context of the historical period in which they occurred, and one could argue that the manner in which we make inferences from test data in the present era is far less scurrilous, the possibility of similar abuses occurring is not as remote as one might suspect (see for example The Bell Curve (Hernstein and Murray 1994)). Nevertheless, ignoring the underlying assumptions of statistical models, or extrapolating data to make inferences beyond the stated measurement envelope of a test instrument can quickly lead us back into a similar quagmire. In an effort to avoid such difficulties, basic testing guidelines have emerged from research and been codified over the last fifty years. It is to these we now turn our attention.

2.2 The development of general testing requirements
By the late 1940's, researchers in the field were suggesting alternative views for determining proper interpretations of test results. They were also encouraging the use of plausible rival hypotheses to explain correlation and get at what Gould called the "underlying essences." (Cronbach 1949) As has been demonstrated above, just because the results of a test correlate with something can not, alone, justify their use. Messick has argued that the way one makes a case for interpreting test scores is "by discovering which multiple lines of evidence support an inference, while making the case that alternative
inferences are less supported." (Messick 1989) Similarly, Shepard asserts that since every test use involves inference or interpretation one must combine empirical evidence and logical argument to support those inferences. (Shepard 1993) Such a definition of how one goes about appropriately interpreting the results of testing also carries some other important implications. First, verifying interpretations arising from data becomes a separate act from verifying a specific test per se. One validates not the test instrument, but the data arising from that instrument. (Cronbach 1971; Messick 1989; Shepard 1993) Virtually any test will yield data with which to make inferences. For example, I test a random sample of the population to ascertain how many of a person’s relatives ate beef during the year 2000 and how many relatives died during that same year. Assuming respondents were truthful in their responses, the data is an accurate representation of test subject’s relatives who both ate beef and died in 2000. Moreover, if this cohort is a random sample of the entire population, the data is also a likely representation of the entire population. However, as noted before, statistically comparing the two responses to infer some hazard posed by beef, in general, is probably not logical. The correlation is probably meaningless. This is not an absolute statement because there could be a reason why such a correlation would suggest a relationship between the two. For example, if in the same study the respondents were also asked if the beef eaten by their relatives came largely from Britain, the correlation between beef and dying might be explained by “mad cow disease.” In both cases, the same test produced the data I used to make inferences and demonstrates, as Cronbach and Meehl point out, that the “principles for making inferences” were validated, not the test itself, since any test is valid for certain purposes, but not for others. (Cronbach and Meehl 1955) Such a belief has been widely held in the field of testing since 1954. (APA 1954)

Along with the belief that one must underpin inferences with both theory and statistical analysis, there arises the idea that multiple measurements better support inferences than does a single measurement. As Messick notes, “One makes the best case for validity by discovering which multiple lines of evidence support an inference, while making the case that alternative inferences are less supported.” To do this, Cook and Campbell argue the need to measure both “convergent” and “discriminate” evidence. The prior confirms that the variables correlate, the latter show they don’t correlate as well with non-germane, intervening variables. (Cook and Campbell 1979) In fact, Messick argues that provisional tests should even contain measures of these non-germane variables (e.g. reading comprehension on a math test) so that one can actually test the correlation statistically. (Messick 1989) Wiggins and others have also noted that the mode of assessment can make significant differences in the data generated by testing media. (Wiggins 1993; Wiggins 1996; Gardner 1997) We often assess learning and understanding using formats that are easiest to score (e.g. the ubiquitous multiple choice test) rather than how well the “behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired domain.” (APA 1985) It has been repeatedly demonstrated that the logic of test construction does not always ensure that a test measures what is intended. We must consider what is “left out in using tests as predictors of future performance.” (Shepard 1993) For example, the research suggests that multiple choice tests favor male test takers, while verbal or open-ended written questions favor female test takers. (NCTPP 1990) In addition, the use of tests
with low or no stakes, versus tests with high stakes, and performance versus pencil and paper tests could produce dramatically different results. (Resnick and Resnick 1990; Wiggins 1996) Loevinger discusses the problems with inferring student ability from pencil and paper tests, but watching in dismay as students who test well are unable to apply the same knowledge in another context. (Loevinger 1957) Anecdotally, teachers, especially in the secondary grades, have also experienced students not taking seriously tests that “don’t count” toward something (e.g. a grade, high school graduation, college entrance, etc.). Wiggins and others also stress the need not only for multiple formats, but also for testing on multiple occasions using different formats, and even integrating testing into normal curricular activities. (Cronbach 1971; Cook and Campbell 1979; Messick 1989; Wiggins 1993)

2.3 The importance of content relatedness
To some extent, the above discussion on how we test presupposes a determination of what we test has already been made. Yet in even the most general tests (such as tests of some supposed “latent” ability like intelligence) determining the actual content domain covered by the tests is critical. Content related evidence demonstrates the degree to which a test is both relevant to and representative of a specific domain of interest. (APA 1985) Relevance requires that a test sample from a specified, bounded domain. While the size of the domain may vary, to provide the data necessary to arrive at valid inferences about a domain the test must only assess content within the specified domain. This forms an upper limit to the breadth of a test. Similarly, a test should also be representative of the domain. After all, it is possible that one may remain within the bounds of a particular domain and still only sample from a very limited part of the domain. For example, if a high school chemistry teacher proposed to test students over the entire domain of high school chemistry by using a test that required students to only balance equations, the test would be relevant, but not representative of the domain. As such the inference that a student scoring well on such a test, knew the entirety of the high school chemistry curriculum well too, would probably be an invalid inference. With the advent of high-stakes tests, even the courts have become involved. In Debra P. v. Turlington 644 F.2d 397 (5th Cir. 1981), the court held that the high school diploma is a property interest and, therefore, subject to the protections of the Fourteenth Amendment. Previously (Lau v. Nichols 414 US 563 (1974)), the court held that the failure of the a school system to provide student with adequate instruction denied them a meaningful opportunity to participate in the public education program and thus violated their rights under the 1964 Civil Rights Act. As a result of these and other cases, states must ensure that tests match the content of curricular materials and what is actually taught. In other words, tests must have content validity. Moreover, students must be notified of the test and consequences, and that the test must not intentionally discriminate against a protected group or class. Section 1019(b) of the Goals 2000: Educate America Act specifically states that, “Assessments developed with funds under title III of this Act may be used for decisions

---

2 Specifically Section 601 of the Civil Rights Act of 1964 states: “No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving federal financial assistance.”
regarding graduation, grade promotion, or retention of students only on the condition that students have been prepared in the content for which the students are being assessed.” Tyler suggests that both skills and knowledge should be evaluated as this would “give a measure of student reactions in a much larger number of situations.” (Tyler 1989)

The methods to determine test relevance and how well the test represents an entire domain often rely on the judgements of “experts,” or an actual performance inventory. (APA 1985; Messick 1989) Generally, these experts first define the domain using a conceptual analysis, then develop test items to represent the intended domain. While random sampling has not worked, certain quantitative methods have been developed to assist this process (e.g. Item Response analysis). However, even these quantitative measures have limitations. For example, discarding items based only on quantitative measures can be problematic. Cronbach and Messick caution against discarding items because they “behave improperly” unless items covering parts of the domain similar to the discarded items replace those previously excluded. (Cronbach 1971; Messick 1989)

As demonstrated in the chemistry example above, overly restricting items almost assuredly insures the domain is not completely sampled and may unfairly penalize a subset of the population. Furthermore, Raizen argues that choosing items that isolate independent facts for testing is especially “antithetical to the process of science” instruction. (Raizen 1990) While non-interdependence of test items is one of the cardinal principals of current test design, the degree of interdependence tolerated among test items is a function of exactly what one is attempting to measure and the inferences data from the test are required to support. All items at their most atomic level are likely to depend on other knowledge for a correct solution. For example, balancing chemical equations or formulas presupposes to some extent that a student has mastered the mathematical concept of multiplication. It also presupposes that the content presented by a test is relevant to a particular audience. While the format in which an item appears can have dramatic, probably unintended, discriminatory effects on sub-populations as was discussed above, how representative test topics relate to the experiences of a culture or group can also significantly alter test scores. (NCTPP 1990; AAUW 1992) Again who we are testing, how we test and for what reasons have profound impacts on the data these tests generate. Using this data, in turn, to support inferences that become student evaluations or educational policy requires that these questions be explored and answered before either making those inferences or settling on a test to generate such data. These observations lead Messick, among others, to conclude that in addition to investigating the content of the tests from which we will draw our data, one must sustain score-based inferences and actions by other evidence. “It is clear that content related evidence cannot stand alone,” he states, “but we need to examine how it functions in concert with construct related evidence in a unified validity framework.” (Messick 1989; Shepard 1993)

2.4 Determining a “construct”

This idea of “construct validity” is the degree to which a measurement device generates data by which we can measure the often hidden traits in a student that account for his or her test performance. It expands the limits of content similarity, as described in the preceding section, to address directly the theory underlying the interpretation of test
results. Messick felt strongly that the questions of content and construct validity were so closely tied that they should be considered as only one facet and not separate facets of test validity. (Messick 1989) His rationale was that content invoked psychological processes other than just content recall in the mind of the person sitting for a test. In addition, Messick felt that validating the content of a test made the test, not the inferences drawn from the data generated by a test, the subject of validation. Yalow and Popham disagreed largely because they felt collapsing the two facets into a single dimension would “compromise attentiveness to content coverage.” (Yalow and Popham 1983)

It seems, however, important to acknowledge a separate role for content validity, even if it is subordinate to construct validity. Evidently, a set of test items could be highly representative of one domain and only moderately representative of another as when, for example, we use a math quiz in math and use the same quiz in physics. The same test would likely have a different level of content validity in the two domains. Linn makes it clear that, “content standards can and should influence the choice of constructs to be measured and the ways in which they are eventually measured.” (Linn 2000) Obviously the content validity does not reside in the test, but in the way the resulting data is used to infer what students know about math or physics. Nevertheless, the data generated by both tests can have important implications for construct validity as well. Perhaps the test is intended to measure some psychological attribute that theoretically spans both math and physics. Consequently, the test device was meant to sample both domains for an overarching attribute. In such a case it might be important to both test the relationship of specific content on the attribute and to demonstrate that the attribute is present regardless of content domain. As in any scientific exploration, the data generated by such exploration both helps support (or dispute) a current theory explaining student performance and also leads to modifications of that theory. In fact Mislevy and others indicate that this process is, by definition, never ending. (Messick 1989; Mislevy, Steinberg et al. 1999)

When considering construct validity, Messick urges one to especially consider two types of construct irrelevant difficulty in test data: Construct irrelevant difficulty occurs when certain non-germane aspects of the test, like reading comprehension problems in a math test, make it more difficult for some individuals or groups. Construct irrelevant easiness, on the other hand, permits some individuals to respond correctly in ways not important to the construct. For example, effective test taking strategies including guessing, time allocation techniques, and using other test questions as prompts to answer a current question. (Messick 1989) Other sources of construct irrelevant data include the mere regurgitation of memorized data or, even more problematic, test questions misinterpreted by the student or answers misinterpreted by the teacher. In the former, when students have been exposed to specific test questions and are then tested on these questions, their responses represent mere recall of algorithms or facts. On the other hand, students who attempt the same problem, but have never seen it before are attempting a much more difficult task. (Bloom 1956; Perkins, Crismond et al. 1995) Attributing such performances to specific constructs such as understanding without supporting data can be entirely unfounded. Similarly, when students misread or misunderstand a question and respond accordingly, the possibly invalid conclusion that may be drawn from the
resulting data is that they don’t possess the construct trait. Likewise, when a test evaluator misinterprets a response, she may attribute more or less of a trait to the test-taker. It must be acknowledged too, that test takers can actively encourage such misinterpretations by cheating, answering questions not asked, or leaving out important, but unknown steps, in open ended answers, etc. To be proposed and tested, a construct must be “embedded in a conceptual framework, no matter how imperfect that framework may be.” (APA 1985) This framework should discuss how the construct relates to other constructs and how measures of the construct relate to other measures.

In this way assessing construct validity parallels Cronbach’s Rival Hypothesis Theory procedure discussed previously. The overlap with content validity questions should be obvious here. The literature suggests that intercorrelations among items that theoretically measure a single construct and the weakness of relationships with items that measure different constructs support the conclusion the construct is, in fact, present. (Cronbach 1971; APA 1985; Messick 1989) The American Psychological Association also suggests the evaluation of individual student responses to verify the existence of a theorized construct, and they suggest evaluating individual performance strategies or responses to particular items to enrich the definition of a construct. (APA 1985)

As before, there are important overlaps between validating this facet of test validity and content validity. Again, both Cronbach and Messick stress the need for repeated testing to allow for “triangulation” on a specific construct. (Cronbach 1971; Messick 1989) Shepard recommends one develop not only correlation statistics, but fully develop prediction equations and to consider what variables are “left out” of the prediction equation. (Shepard 1993) Anastasi calls construct validity the “superordinate category” that can only be evaluated through logic checked against real life criteria. (Anastasi 1986)

2.5 Specifying the “criterion”

Criterion validity describes how well test scores compare to some external measure of the behavior in question. “The fundamental question is always: ‘How accurately can criterion performance be predicted from scores on the test?’” (APA 1985) Criterion validity can be further divided into two sub-categories: short term (concurrent) and long-term prediction of behavior against a criterion. Here again, the modality of the test that generates data is of particular import. One does not test driving skills, for example, with a written test. On the other hand, it might be just as inappropriate to test abstract symbolic manipulation with concrete manipulatives. Howe states emphatically that the fact “performance assessment is potentially more predictively [sic] valid than traditional measures, is unassailable, but the way we weight [such assessments] could be more problematic.” (Howe 1995) Howe further describes two types of problems with the manner in which the criteria themselves are constructed. Within-criteria bias occurs when the criteria itself discriminates against a group for no justifiable reason. An example of this was the exclusion of women from US military service academies prior to 1980. Across-criteria bias occurs when the weighting of multiple criteria into a single standard unnecessarily penalizes certain individuals or groups. An example cited by Howe is the raising of cutoff or passing scores when positions are scarce. (Howe 1995)
Using arbitrarily set cutoff scores as a criterion can be particularly problematic. To see why this is true and the difficulty of using test results to make such decisions consider Figure 2-1. In this figure, the oval represents the criterion validity of a test in that it delineates the correlation between a test score (the horizontal axis) and the probability the respondent could, in fact, perform the tested criterion satisfactorily (the vertical axis). In Figure 2-1 an arbitrary correlation between these two variables of $r = .35$ is assumed. Test takers in the upper right hand quadrant both successfully pass the test and successfully perform against the stated criteria. Similarly, test takers in the lower left hand corner are those who fail the test and perform unsatisfactorily against a criterion. Test takers in the upper left quadrant and the lower right quadrant have been misclassified. The former group are false negatives – those who would have performed acceptably against the criterion, but failed the test. The later group (the false positives) are those who passed the test but perform inadequately against the criterion.

![Figure 2-1](image)

**Figure 2-1.** When test results are used to make decisions, subjects are divided into four groups based on the quadrants formed by two orthogonal lines. The vertical line represents the division between passing and failing a placement test. The point where this axis crosses the horizontal axis represents a threshold ("cut" score) that students must meet or exceed to pass the test. In this figure, the threshold is set at the 20th percentile (80 percent of the students pass). The horizontal line represents a theoretical division between those students who actually meet the desired criteria (above the line) and those who do not (below the line). Students in the upper right and the lower left quadrants are correctly classified. However, students in the upper left quadrant (false negatives) fail the test but could have performed satisfactorily, and students in the lower right (false positives) pass the test but should have failed. The oval represents the actual dispersion of the population when student percentile is plotted against probability of meeting or failing to meet the criteria. In this case, the correlation between these variables is $r = .35$.

Where one sets the threshold (i.e. the point where the vertical axis crosses the horizontal axis) ultimately determines how large each of these groups will be. As the vertical axis is moved to the right and a higher percentile score is required to pass, both the number passing who can meet the criterion and the number of false positives decrease. The number who fail the test but could have performed adequately, increases. Conversely, when the threshold is moved toward lower percentile threshold values, those passing and those who pass but cannot meet the criteria (i.e. false positives) both increase. For
example in Figure 2-1, if the cutoff score were set at the 20th percentile (80% of test takers pass), about 17% of those who pass would be unable to perform against the criterion. Approximately 66% of those who fail, however, would have performed successfully had they been admitted. (Messick 1989; NCTPP 1990) While placing the bar is a value judgment made separate from test construction, it is apparent that tests generating data correlating more strongly with the criteria measure (e.g. r = .9) significantly reduce the number of false positives and false negatives. As the correlation coefficient (r) approaches the value of one, the ellipse would collapse to a line and the total number of false positives and negatives would fall to zero.

The National Research Council has found that group differences in test performance were consistent with broad social inequities (e.g. family income, educational opportunities, and other economic disadvantages) rather than actual performances in education or on the job. Therefore, because underrepresented minorities do less well on present tests, they tend to be misclassified at a higher rate, regardless of the reason for misclassification, if such classifications are based on tests alone. Since such quantitative classifications are “easy to do and our society is dazed by quantifiable measures we should not assume such classifications are unusual.” (NCTPP 1990) In fact, the NCTPP cites an example of a 1976 version of the Armed Services Vocational Aptitude Battery (ASVAB) that was miscalibrated. More than 300,000 recruits who would normally have been rejected were admitted into the military. Studies showed that many of these enlistees fared as well as or better than their higher testing peers. As a group, they performed only slightly less well than the other cohort. (NCTPP 1990)

Furthermore, the National Council on Testing and Public Policy (NCTPP) finds demonstrable evidence that alternatives to the ubiquitous multiple choice test are significantly less biased against ethnic, cultural and language minorities. In fact, the NCTPP argues that when test scores are used to disproportionately deny opportunities to minority groups, there is a greater need to show that tests actually measure the characteristics being evaluated. (NCTPP 1990) Howe goes further in suggesting that “a test is biased” if it under-predicts any group’s performance on the criterion of interest relative to some other group. (Howe 1995) However, this seems somewhat extreme. Again it is important that the examination of criteria validity fall back on underlying theory. For example, the fact that female students typically earn higher grades in secondary science classrooms and that classroom tests correlate strongly with that grade measure should not automatically indicate bias. At least part of the correlation might be easily explained by the fact that, on the whole, female students are more consistent in accomplishing homework than are their male peers. While the better test scores may be disproportionately awarded to females, such a statistic prompts further investigation into the cause of this apparent inequity not an immediate conclusion of bias.

The need for underlying theory is also important for another reason. Norris, for example, argues that because IQ correlates with reading achievement better than “just about any other test” IQ should be used to measure reading achievement. (Norris 1995) This parallels the argument made above, and should be discounted for similar reasons. Unless
there is underlying theory or logic to explain such a correlation, the validity of such practices should be considered suspect.

Obviously, the context of the initial test performance is also of particular importance as varying contexts can dramatically alter performance. For example, environmental variables such as temperature, or test subject fatigue can have adverse effects. If such variables are dramatically different between initial and subsequent performance, similarities between performances could easily be different and the predictability of the test minimized. The motivation of students to accept or achieve a certain criteria must also be recognized, especially in the secondary and post-secondary schools, and particularly on low / no stakes tests. We need, as Thorndike stresses, to evaluate if the test generates the data that relate to the ultimate outcome sought by the decision maker. (Thorndike 1922). As before, the need to use multiple measures to generate data seems crucial and unavoidable. (Messick 1989)

2.6 The “Standards”

The American Psychological Association (APA), along with the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) have formally outlined the requirements developed in the literature cited above as a series of Standards for Educational and Psychological Testing. The purpose of this document is to provide a comprehensive basis for evaluating tests. This process began in 1956 and revision continues to the present. The Standards (as they are more commonly known), refer to test validity as “the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences from the test scores. Test validation is the process of accumulating test evidence to support such inferences.” They add that, “well-constructed and validated tests provide a better basis for making some important decisions about individuals and programs than would otherwise be available.” (APA 1985) Logically then, tests that generate data which produce inferences that are invalid in particular or in a number of instances can have devastating effects on individuals and policy decisions. Moreover, these inferences can actually be counter-productive to assessing student learning and understanding, informing curriculum development and evaluating our educational institutions. The Standards assign the test user the responsibility of accurately describing the extent to which the test score data support particular decisions or inferences and to provide evidence of validity of the data for a particular decision or assessment (my emphasis). (APA 1985) Although one cannot ignore the responsibility of test developers to define and evaluate the parameters within which a test generates the data on which to base inferences, the Standards make it the responsibility of the test user to stay within the prescribed parameters and to restrict inferences to those supported by the data.

The standards make some important recommendations concerning testing that are especially relevant to the present discussion. As much as possible I will group these according to the general categories of validity discussed previously.
2.6.1 Assuring Content appropriateness

♦ A clear definition of the universe represented, its relevance to the proposed test use, and the procedures followed in generating test content to represent that universe should be described. (Standard 1.6)

♦ Test taker’s scores should not be accepted as a reflection of lack of ability with respect to the characteristic being tested for without consideration of alternate explanations for the test taker’s inability to perform on that test at that time. Possible explanations involve linguistic or cultural backgrounds. (Standards 6.10 and 6.11)

♦ When a test is used to make decisions about promotion or graduation, there should be evidence that the test covers only what student have had an opportunity to learn. (Standard 8.7)

♦ Students should have multiple opportunities to demonstrate skills when promotion or graduation is at stake. (Standard 8.8)

2.6.2 Assuring Construct appropriateness

♦ When a test is proposed to measure a construct, evidence should be presented to support such an inference and should show that the test does not depend heavily on extraneous constructs. It should also show that the score is more closely related to that construct than to other, substantially different constructs. (Standard 1.8 and 1.9)

♦ It should not be assumed that because a test title contains the term “aptitude” or “ability” that the test measures a construct distinct from what is measured by an “achievement” test. (Standard 8.3)

2.6.3 Assuring Criterion appropriateness

♦ Criterion related evidence is the most common approach to validation in admissions contexts, although both content-related and construct-related evidence are important. (Preface to Section 8 – Educational testing)

♦ Using test scores for classification or certification decisions presupposes a consensual body of material upon which all students can be reasonably compared. This implies not just standard textbooks, but systematic instruction and a minimum period of study. (Preface to Section 8 – Educational testing)

♦ A criterion-related validation study should provide a description of the sample and statistical analysis used to determine the degree of predictive accuracy. (Standard 1.11)

♦ A test used to certify that a student should be promoted or graduate (i.e. a gate-keeping or accountability test) should describe the instructional and test domains in
sufficient detail to allow evaluation of the agreement between test domain and content domain. (Standard 8.4)

- Decisions that will have a major impact on a test-taker should not be made on the basis of a single test score. (Standard 8.12)

2.6.4 Assuring Analytical appropriateness

- Test users are responsible for validating the inferences from a test used for a purpose(s) that has (have) not been previously validated. (Standard 6.3)

- Test users should remember that inferences that are based on an individual’s response to single items are subject to considerable error. (Standard 1.4)

- Test users should act to counter the tendency of people to attach unsupported surplus meanings to test scores. (Standard 6.5)

- Test users should verify that changes in available techniques have not made more dated procedures inappropriate. (Standard 6.7)

2.7 Current (representative) assessment instruments and techniques

As discussed above, the evolution of procedures for developing, administering, evaluating and interpreting tests has proceeded in earnest during the last century in America. Often these procedures arose as a result of unsupported or wholly unwarranted inferences made from the data produced by the test itself. In part, procedures changed when the implicit or explicit theory underlying the data interpretation was exposed as illogical or untenable (e.g. the use of sensitivity to a whistle or head size to explain intelligence). As the century progressed, statistical methods gained increasing popularity as a means to identify problematic theory. Nevertheless, questioning a theory or current practice because it seems to fly in the face of experience has also been an important source of new practice and procedure in the field. This seems especially important in the area of content validity where specific analytical methods have not yet been developed. As in other fields of endeavor, new theories and procedures replace the old in order to better explain events or objects in the world. In addition, new theories arise as new measurement variables are proposed and as changes in technology allow new measurement techniques. The use of Positron Emission Tomography scans to investigate the neural basis of Spearman’s g factor (Duncan, Seitz et al. 2000) is a good example of such a technology. This dissertation arises from such technological advances.

The theories of what constitutes learning, understanding or cognitive development seem, at present, to be in just such a state of change and expansion. (McDonald; Gardner 1983; USDoEd 1983; Bloom 1984; Sizer 1984; Resnick 1987; Schwartz 1989; Raizen 1990; Wolf, Bixby et al. 1991; Herman, Aschbacher et al. 1992; Papert 1993; USDoEd 1993; Wiggins 1993; Maker 1994; Siklossy 1994; Wisner 1994; Newmann, Secada et al. 1995;

There is great uncertainty about what constitutes educational effectiveness in large part because there is great disparity in what the citizenry, policy makers, and researchers recognize as educational achievement. Some believe choosing the correct answer from a group of possible answers is enough to demonstrate learning. Others go further. They believe students should demonstrate the ability to arrive at a correct answer on their own. Allowing students to generate an answer and “following” their work as they do so also offers the opportunity to gauge the degree to which a student can generate an answer. In both cases, proponents of these positions share an underlying belief that ability on the test predicts future ability to use the tested knowledge or skill regardless of context. During the last forty years, still others in America have defined student achievement in terms of how well students can apply concepts to real-world problems. In addition some have proposed measuring variables like drop-out or matriculation rates, attendance rates, student motivation, propensity to be a life long learner, and ability to work with others.

In general, the previous paragraph only dealt with the way data is generated. Once developed the data can be analyzed in various ways to satisfy the wants and need of consumers. For example, test takers can be ranked relative to one another or, if the test was calibrated to a statistically “normal” population, ranked in a variety of ways (e.g. percentile, grade equivalent, stanines, standard deviations, etc.) against the norm. For this reason these techniques are termed normative assessment. When deciding whether a certain norm is appropriate for comparison one should consider the demographic similarities of the “normal” and tested group, when the test was last normalized, and the size of the “normal” group. An alternative to normative assessment is to compare students to a set of standards or criteria to assess achievement. As Wiggins indicates such assessments were very common in the schooling of apprentices. (Wiggins 1993) For a host of reasons explored below, there appears to be a growing number of advocates for these types of assessments in present day school settings. The term authentic assessment has been used to describe this “criterion referenced assessment” when it involves students solving “real-world” problems.

A third dimension in how testing data will be used involves the uses of the test inferences. These uses typically fall into two categories: formative and summative. In formative applications of test inferences, the main purpose of testing is feedback, usually to the classroom teacher or other curriculum designer. It can also help student and teacher correct misunderstandings or misconceptions in understanding and reorient the student or the educational goal the student is attempting to achieve. Summative assessment, on the other hand, seeks to evaluate. In this sense, the use of test data is more a measure of accountability of student, teacher, or school.

While the line between summative and formative, or normed and criterion referenced testing is not absolute, care must be taken when using data from one type of test to make
inferences requiring data from the other. Popham explains that, in most cases, summative tests that are standardized and normalized (achievement tests) are most often designed to rank order, that is discriminate, one student from another. Therefore, the questions these tests often contain are the questions that best do that. While content is inevitably considered in developing these tests, the huge size of the content domain coupled with the need to build a test that spreads students along the normal curve, means that a question's discriminating power is ultimately more important than content coverage. Moreover, if a student's academic achievement is excellent compared to the standards under which s)he is taught, but a test used to evaluate the student's performance is normalized to a student group evaluated against different standards, the data can only indicate how well the student did against those other students, not how well s)he did against the appropriate standards. (Popham 2000) Unfortunately, Freeman and colleagues found that only 20 – 50% of what was on standardized achievement tests were suitably addressed in text books. (Freeman, Kuhs et al. 1984)

I now apply the concepts of inference validity to six measures which typically inform the American public and policy makers about student achievement. The inferences these stakeholders make about school effectiveness (accountability), student achievement (certification / aptitude (ability), or promotions (placement) are often the result of using data generated by these tests. As discussed previously, if the inferences are not valid, improper consequences are likely to result. For example, teachers may be fired, students misclassified, or students not promoted, graduated or admitted to the next level of education. As we shall see, what we need to make the inferences required by present standards cannot be supplied by these tests alone as they are presently constituted and administered.

First I will consider the inferences resulting from the Third International math and science Study (TIMMS). Second, I discuss the National Assessment of Educational Progress (NAEP) and the SAT as representative of national tests. Third, I look at the Stanford Achievement Test – ninth edition (SAT-9) as applied in the state of California and other states. Finally, I address the inferences made about classroom grades as measures of student achievement. In each case, my goal is to briefly describe some of the valid inferences data from these measures support and to highlight the major inappropriate inferences arising from each. I will close this chapter with a discussion of authentic performance assessments.

2.7.1 The Third International math and science Study (TIMSS)
The Third International mathematics and science Study (TIMSS) was conducted in 45 countries around the world in 1995 to measure student achievement in math and science at the 4th, 8th and last year of secondary school grade levels. It was the largest such comparative study to date and consisted mostly of written response items. Individual countries also had an option of administering a performance assessment. Because of time and money costs, students participating in performance assessments were not observed directly by raters, but recorded answers in workbooks and submitted their workbook and product for later scoring. (TIMSS International Study Center 1995) In the United States
500 schools and approximately 33,000 students participated at all three grade levels of the TIMSS. However, this level of participation was unusual as fewer than 20 nations participated in all three TIMSS levels. In addition to comparing student achievement, the TIMSS was also intended to allow nations to compare their national curriculum, and to measure the effects of content coverage on student achievement. This was done through the use of questionnaires (Jakwerth, Bianchi et al. 1997) The TIMSS achievement tests were developed through an international consensus-building process involving input from international experts in mathematics, science, and measurement. Furthermore, the tests were endorsed by all participating countries. (TIMSS International Study Center 1995)

As Elmore notes, the structure of the data resulting from the TIMSS is complex. (Elmore 1997) Not only did various countries participate to varying degrees (raising the possibility of selection bias), but the potential variation in content coverage is enormous. The former suggests that one cannot make comparisons across test populations that are not common. For example the National Center for Educational Statistics (NCES) states, “The U.S. students scored above the international average in both mathematics and science at the fourth-grade level. At the eighth-grade level, U.S. students performed above the international average in science and below the international average in mathematics. U.S. twelfth graders, including our most advanced students, performed among the lowest in both science and mathematics.” (NCES 1999) Such a statement is misleading. In this case the population taking each test is dramatically different, so there is no “international average” per se. In addition, Holliday reports that only five participating nations claimed to adhere to the promulgated sampling standards. (Holliday 1999) Not only is the specter of selection bias ignored, but even more critically the statement implies that the fourth-grade curriculum in the U.S. is fine, the eight-grade curriculum is slipping in math, and that secondary education in the United States needs a lot of work to “catch-up.” In fact, student performance in one grade is the cumulative result of thousands of hours of experience in and outside the classroom. Such experiences are also dependent on socioeconomic status, race, and home environment. (Elmore 1997) The TIMSS methodology does not generate the data to support these inferences even though the NCES suggests they do. An even more ominous problem with using TIMSS data to make inferences about the learning and understanding of American students arises because of content validity questions.

The makers of the TIMSS created tests to compare student learning around the world. Unfortunately, there is no standardized world-wide curriculum. Jakwerth, for example, demonstrate that there is wide variability in math curriculum for 13 year olds both across and within nations. Not surprisingly they find that the content of the TIMSS is more similar to the course content of some countries than it is to others. Furthermore, these researchers cite no fewer than ten studies indicating that the “opportunity to learn the skills being tested is a significant explanatory variable of student performance.” (Jakwerth, Bianchi et al. 1997) Berliner concludes that such differences are probably explained more by differences in various curricula than by the effectiveness of a particular national system of education. (Berliner 1993)
The validity of inferences about student abilities in the United States, as a whole, based on the TIMSS data is probably invalid for another reason. The makers of the TIMSS built the instrument on the stated or the "intended" national curriculum if a national curriculum did not exist. Developers uncovered such an "intended" curriculum in textbooks, among other sources. Unfortunately, in the United States, textbooks are often written at a national level to cover a wide audience and so address the different standards of many states. As such the content can be quite diverse, but any single aspect of that content may only be implemented very sporadically. Furthermore, given the decentralized nature of the American educational system, the curriculum that is actually implemented might vary dramatically from classroom to classroom within a state. In the case of the United States, it may be more valid to make inferences and comparisons between states and other nations than to make those comparisons between the United States and other nations.

A final area of concern regarding TIMSS is answer rating. The nature of what constitutes a “right” or acceptable answer and how this might skew resulting data has been a concern at least since the days of the first intelligence tests in America as demonstrated above. This is especially true when sampling across cultural boundaries. Messick notes that "standardization of test materials and administration conditions could actually be a source of invalidity across cultural settings." He continues that, "individuals from different cultural backgrounds might display their skills or engage task processes more optimally if the task conditions were altered to be more in line with the respondent’s cultural context.” (Messick 1989) This was not always the case with the TIMSS. So, for example, if an U.S. eighth grader was unaware of what “cm” (centimeter) meant on a TIMSS math test, they could have provided incorrect answers even though they knew the mathematical concept being tested. Other examples include the need to take five or more readings of pulse (an apparently arbitrary number) for full credit on a science performance assessment when an exact number of readings was not specified in the problem statement (see rubric for “Pulse” problem). (TIMSS International Study Center 1995)

As an aside, the effects of these tests on educational policy and how they affect pedagogy must be considered. Messick believes these consequences can have detrimental effects on the usefulness of inferences made from test data and, therefore, should be examined as another aspect of validity. (Messick 1989) While others disagree (Popham 2000), what researchers seem to agree on is that the implications of the test on education should be considered from a policy standpoint. Resnick, for example, notes that sampling independent bits of knowledge and skills remains the basic strategy for standardized testing. In fact she argues that the usual multiple choice format of standardized testing reinforces the idea that the student’s task is to find or guess the right answer, not to explain their reasoning about how they arrived at an answer. (Resnick and Resnick 1990) Atkin reaches similar conclusions, “External assessments in particular...have traditionally exerted powerful control over what teachers feel obliged to do. Yet the impact of such assessments on efforts to engage students in original and complex work can be devastating if the tests measure only memory.” (Atkin and Black 1997)
While inferences about the effectiveness of various national curricula to achieve success on the TIMMS might be an appropriate inference from the TIMSS data, inferences about student ability or the comparison of one nation's curriculum or students to those of others is inappropriate. The questionable inferences drawn from the TIMSS and its possible negative consequences for US education policy make its value for assessing student learning and understanding suspect.

2.7.2 The National Assessment of Educational Progress (NAEP)

The NAEP (also known as the “nation’s report card”) is a Congressionally mandated project overseen by the National Center for Educational Statistics (NCES) at the US Department of Education. The NAEP is designed to be an assessment of what America's students know and can do in various subject areas at the national level. The NAEP has collected two forms of data since its inception in 1969. The long-term trend assessment form of the NEAP is administered every two years in a form identical to the 1969 version of the NEAP. On the other hand, the national NAEP has, for the most part, been administered to fourth, eighth and twelfth graders every year since its inception in 1969, but has undergone various modifications to address changing curricular, pedagogical, and political requirements. As the national NAEP samples were not designed to support the reporting of accurate and representative state-level results, the Congress allowed State versions of the NEAP at the fourth and eighth grade level beginning in 1990. The national and state forms of the NAEP assess various curricular areas in different years and use a matrix form of assessment in order to minimize impacts on student instructional time. As such the NAEP is broken up into a dozen or more blocks, consisting of multiple items, and only two or three blocks of items are administered to any particular student. This methodology, therefore, means that the performance of any particular individual student cannot be measured accurately. In fact, by law NCES is prohibited from reporting the NAEP results of individual students. (NCES 2000)

The curricular subjects each version of the NAEP assessed in 2000 are shown in the figure below:
NAEP 1999 and 2000 Assessments

1999 NAEP

<table>
<thead>
<tr>
<th>Long-Term Trend</th>
<th>Age 9</th>
<th>Age 13</th>
<th>Age 17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mathematics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Science</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th>Grade 8</th>
<th>Grade 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

2000 NAEP

<table>
<thead>
<tr>
<th>Main NAEP</th>
<th>Grade 4</th>
<th>Grade 8</th>
<th>Grade 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Science</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State NAEP</th>
<th>Grade 4</th>
<th>Grade 8</th>
<th>Grade 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2-2. This figure lists the content areas addressed by the three forms of the NAEP in the year 2000. Although math and science are tested in the year 2000, the curricular domain rotates from year to year. In 2001, History and Geography will be assessed.

(Source: http://nces.ed.gov/nationsreportcard/guide/ques2.shtml)

NAEP assessments use multiple choice, constructed-response questions wherein students must write responses ranging from two or three sentences to a few paragraphs, and performance tasks that require the use of calculators and other materials.

As might be expected, there are two areas of concern when deciding what sorts of inferences are supported by NEAP data. First, how are students selected to participate in the NEAP and how is selection bias avoided? Second, how closely does the content tested by NEAP match the material taught in the nation's classrooms?

Student participants are selected by NAEP staff based on the school they attend. When a sample school has been chosen, all students in a targeted grade have an equal probability of being asked to take the NAEP. Students are encouraged to take the test but may opt not to do so. The NCES acknowledges that validity problems can arise, however, if selected schools decide not to participate, if selected students are absent or refuse to participate, or if schools or students volunteer to join the NAEP sample. For this reason, NCES encourages the participation of all those selected, but does not allow schools or students to volunteer for the test.
The National Assessment Governing Board (NAGB) develops an organizing framework for each subject to specify the content that will be assessed. It is important to note that the framework for each subject area is determined through a consensus process that involves teachers, curriculum specialists, subject-matter specialists, school administrators, parents, and members of the general public at the federal level and so does not explicitly conform to particular state or local standards. Under the direction of Educational Testing Service (ETS), teachers, subject-matter specialists, and measurement experts develop the actual questions and tasks based on the frameworks. Subject matter experts then review the questions ETS has generated in each content area.

As was the case for the TIMSS, questions about the NAEP's content validity are inevitable given the methodology just described. These questions have been addressed in two ways. First, the NCES has commissioned various studies of content validity. I, however, only find studies addressing how closely actual NAEP questions respond to the frameworks. For example, while Sireci, et al. conclude that, "the majority of the [NAEP] items studied (85 percent) were judged to be measuring the content areas they were designed to measure" (Sireci, Robin et al. 2000), the "content area areas they were designed to measure" are those specified by the National Governing Board frameworks, not state or local content standards. Although such items "test what they say they do", as demonstrated for the TIMSS above, one should have serious concerns if students are tested over material they have not had an opportunity to learn. In fact, Muthen, et al. found that different curricular content factors at both the 8th and 12th grade did explain differential NAEP performance. (Muthen, Huang et al. 1995) The lack of content validity studies should be especially worrisome to those who intend to use NAEP data to make judgements on the effectiveness of teachers or schools within specific states or districts. At times the differences between the NAEP and other tests can be dramatic, as happened in the state of Louisiana in 1997. The Louisiana state test showed 88% of its students reading at the "proficient" level, while the NAEP indicated only about 15% of Louisiana's student were proficient readers. (Sacramento Bee 1997) Louisiana is not alone. Smith, et al. cites a Southern Regional Education Board Study that reveals similar findings in other states. (Smith, Stevenson et al. 1998) In fact, in a study conducted for the U.S. Congress, the National Research Council concluded that state and commercial tests vary too much in content, format, and sampling techniques (among others), to allow effective student comparison. (Fox 1998) Linn also cited the difference in purpose between the two tests (i.e. the criterion). In Wisconsin, 88% of students met or exceeded the state standard for reading at better than a marginal level. Only 35% of these students, however, met the NEAP standard of "solid performance." Suggesting that these two standards require the same level of competency is at best misleading, at worst ludicrous. (Linn 2000)

Another question about the appropriateness of inferences made from NAEP data has also been studied. As I have previously mentioned, the NAEP is an assessment system designed to have minimal impact on the test-taker. Given its no-stakes/low-stakes nature, the impact of test-taker's motivation on the results is an inevitable concern. As older children tend to be more savvy in their understanding of test "stakes", studies of
motivational effects on NAEP scores focus primarily on eighth and twelfth graders. Shanker and my own classroom experience suggest that “if students know that what they do on a test doesn’t matter, they may decide it’s not worth their while to put forth any effort. And it could be that this explains the low level of achievement we have seen on NAEP examinations.” (Shanker 1990) Linn, Koretz, and Baker report on a study that embedded NAEP mathematics items in a state accountability assessment and found a small, statistically significant effect on problem performance of 8th graders. (Linn, Koretz et al. 1996) Given that the test these questions were placed into was of relatively low stakes, perhaps such a small improvement is all that should be expected. O’Neil, et al. found similar effects in eighth graders who were motivated to solve NAEP problems by financial rewards. They also noted that while twelfth graders did not show improved results when offered financial incentives for better scores, they did report trying harder. The overall conclusion from this study was that trying harder only increases test scores when increased investment of effort permits greater retrieval and use of the relevant prior knowledge one possesses. “It does not affect performance when prior knowledge is weak.” Might a rival hypothesis be that these 12th graders had not seen the content being tested in their curriculum and so they lacked the ability to turn motivation into success? Perhaps, but the overall conclusion seems clear; like the TIMMS, the NAEP has its appropriate and inappropriate uses. Evaluations of individual student achievement and school effectiveness are not two of the appropriate uses.

While the inferences drawn from NAEP data seem plausible to support general conclusions about trends in student attainment of specific national goals, the data does not allow similar conclusions at a state, district or school level. Even the inferences about national trends might be questionable until more research is conducted on the effect of student motivation when sitting for no- or low-stakes testing. Nevertheless, the lack of demonstrated content overlap between NAEP and the local curriculum, and the known diversity of those various curricula between schools suggests that, like the TIMSS, data from NAEP cannot support accurate inferences about individual student achievement.

2.7.3 The SAT

The SAT (formerly known as the Scholastic Aptitude Test, but now just known by the initials SAT) (PBS 1999) is a national test administered by the non-profit Educational Testing Service (ETS). Its express purpose is to identify students who will be successful in college. (Shepard 1993) Because of this explicit connection to college admission and the fact that most American colleges require it of applicants, the SAT is considered a high-stakes test. The SAT traces its roots to the intelligence tests developed in the United States in the early 1900’s and an altruistic desire on the part of the then-president of Harvard University, James Conant, to admit students based on merit. Candidates with an “aptitude” for higher education versus achievement were sought. (PBS 1999) In fact the SAT and IQ test results correlate highly. Although, because of the name change, a connection with a “latent” trait such as aptitude or intelligence is no longer implied, such a link still forms the basis of SAT construction. Wayne Camara, head of the College Board’s Office of research, maintains it tests “developed reasoning skills” linked to the breadth and depth of the school curriculum a student has been exposed to as well as “out of school learning.” (PBS 1999) Nevertheless, the organization FairTest and others
maintain the test actually judges quick strategic guessing with less than perfect information. In fact, Guinier cites interviews with test takers that suggest those who “work through problems” get lower scores on the SAT than those who speed through the problems and guess at answers. (PBS 1999)

The SAT is designed to rank students on a scale of how well they will do their first year in college, not to determine how well they have learned or can apply concepts taught in school. The SAT is not designed to give teacher or student feedback. For this reason, the SAT, unlike the TIMSS or NEAP, is classified as a normative, aptitude test, not an achievement test. In selecting questions for the SAT, attention is given to questions that will “discriminate” students. As such, the content these students have been exposed to is of secondary consideration. Jerard Kehoe at Virginia Polytechnic Institute and State University explains that such “item analysis” improves multiple-choice tests by evaluating the extent to which questions are able to measure a single skill, improve the effectiveness of “distractors”, and eliminate items that all students answer correctly or incorrectly. Kehoe recognizes that some may question the ethics of throwing out poorly discriminating questions, especially because such an action explicitly seeks to eliminate the very items less able students tend to do well on, but that more able students tend to miss. He, nevertheless, justifies the practice by reminding his reader that, “the purpose of testing is to determine each student’s rank.” (Kehoe 1995)

But what does ranking students really tell us about what a student has learned or can do? The answer is not always a lot. Take for example the following situation. A class of students is tested and ranked according to a normalized group of students. Assume that in this first case, within the normalized group, students above the mean all meet or surpass a minimum competency threshold, while those below the mean do not. That situation is represented graphically in Figure 2-3 below.

![Figure 2-3](image)

**Figure 2-3.** This figure shows student test scores as a normal distribution. In this case a mean score on the test corresponds to the minimum score determined to demonstrate competency. Furthermore, approximately half the students meet minimum competency requirements and half the students do not.
Now assume that another group of students is tested and their scores are also normalized to the original group, but that the competency standard for this second group is set much higher than those of the first group. In this case, even though it appears students ranking high in the second group (because they scored higher than the mean) have done well, none have met the minimum competency requirement established for them. See Figure 2-4.

![Diagram of a normal distribution with a minimum competency score mark]

Minimum Competency Score

Figure 2-4. This figure again shows student test scores as a normal distribution. However, in this case the minimum score for competency is set well above the mean score on the test. Even though half the students score above the mean on the test, few of the students (perhaps none), demonstrate what is considered to be minimum competency.

This simple example should demonstrate the folly of attempting to use scaled or normalized scores alone to determine competency or student achievement without tying such scores to some sort of meaningful criteria. (See (Messick 1989)) In fact, Guskey makes the statement that evaluation on “the curve” makes learning a highly competitive activity in which students compete against one another for a few scarce rewards. (Guskey 1996) But this type of use of the SAT is precisely what Powell and Steelman document in their report *Bewitched, bothered, and bewildering: The use and misuse of SAT and ACT scores*. In fact Powell and Steelman describe how ways to “beat” the ranking systems have become common. (Powell and Steelman 1996) It should also be noted that normalized scores not only might give misleading results, but that by definition such a procedure commits half the population to scoring below the mean in each normalized sample. This fact can be lost on policy makers. For example, some “districts are currently being encouraged to set as a target for Title I funding, that 90% of their students score above the national median on a standardized test within the next 10 years.” (Linn 2000) In fact, Claude Steele has observed that even when tests are normalized, scores are difficult to interpret if they are not at one of the extremes. In the case of the SAT, for example, he observes that a student who received anywhere from 10 to 1200 may have obtained the scores because of coaching, or a lot of family travel, or because s)he happened to know what was on the test that day. (PBS 1999) The fact that the SAT is
time constrained also seems to contribute to its “coachability” as students who know what to expect need not spend valuable time reading directions or examples. In fact, John Katzman, President of the Princeton Review\(^3\), cites his experience that students do not do well on the SAT not because they get hard questions wrong, but because they make many “careless mistakes” on easy questions such as choosing answers intended to distract students from the correct answer. (PBS 1999) ETS has, on the other hand, countered that the SAT cannot be coached. Nevertheless, a meta-analysis conducted by Kulik, et al. found strong evidence that coaching does result in small positive gains for takers of the SAT and that no significant negative effects from coaching have ever been reported. (Kulik, Bangert-Drowns et al. 1984) These findings should make us take pause, for as the Standards council, “An interpretation that says that a test measures abilities that are developed over the course of twenty years and that those abilities change slowly as the result of time and effort would be called into question by evidence of significant changes that result from short-term coaching.” (APA 1985)

In its defense however, one must recall that the ETS has not overtly promoted the use of the SAT as a measure of student achievement. I have found nothing in the available literature indicating that ETS advocates the use of the SAT to compare educational effectiveness of schools, districts or states. This role has been thrust on the instrument by others. Wiggins argues that “although not designed to be an achievement [test, the SAT] is used to whip districts and states into a frenzy about school achievement – despite explicit warnings by ETS not to do so and despite the obvious fact that such tests depend largely on aptitude associated with socially accepted and narrow dimensions of intelligence (and, thus, socioeconomic status).” (Wiggins 1990) ETS advocates one use for scores generated by the SAT: the prediction of freshman college grades. Generally the available literature agrees that the SAT, in fact, predicts approximately 18% of the variance in freshman college grades (Crouse and Trusheim 1988; PBS 1999), although early in 2000, an official at the ETS estimated it was only between 10 and 15% of the variance. (Drake University 2000) In fact, Gardner finds the predictive ability of “scholastic aptitude tests is often questionable in view of their limited usefulness in predicting performances beyond the next year of schooling.” (Gardner 1997)

Nevertheless, the ability of the SAT to explain some of the variance in freshman grades seems relatively consistent among both highly selective and non-selective colleges. (Crouse and Trusheim 1988) The College Board reports that the reliability of the SAT is high (r= .9 between tests (College Board 2000)); however, it does not report for what score range or tolerance this figure is calculated. In addition, even at this level of reliability, test scores can vary dramatically from the student’s “true” score or the score they would receive if they took the test again. (Rogosa 1999) Even though the correlation between the SAT and freshman college grades is statistically significant, other problems suggest that using them for this purpose can lead to inappropriate inferences. Typically, SAT scores “under-predict women’s grades and over-predict men’s. Young women tend to receive higher college grades than young men with the same SAT scores.” (AAUW 1992) In addition, the SAT does not appreciably improve college grade forecasts of either black or low-income applicants. (Crouse and Trusheim 1988) Arguably more than just

---

\(^3\) The Princeton Review is an SAT coaching company which prepares 100,000 students annually to take the SAT.
SAT scores account for these differences. While there is a correlation between family income and test scores, other indicators of educational achievement, including school grades, have similar relationships to students’ economic backgrounds. “These are reflections of a fact that, in our society, students from higher income families enjoy educational advantages that many lower income students do not.” (ETS 1980)

Clearly the above discussion suggests that use of SAT data to make certain inferences should be limited. Specifically, using SAT scores to support absolute inferences about student achievement is clearly inappropriate, as is the use of these scores to make inferences about the effectiveness of educational curriculum or institutions. Furthermore, as some apparently serious biases exist, care should be taken when using the SAT to predict the future performance of women or underrepresented minorities. As ETS recommends, when used, the SAT should only be one of many data points upon which to base decisions like college admission.

2.7.4. The Stanford Achievement Test – 9th Edition (SAT-9)

The SAT-9 is an “off the shelf” student achievement test used in California and other states as part of a state level, high stakes student accountability system. In some respects, the SAT-9 resembles the NAEP. First, the test is normalized to a national sample of students. The scores reported from the SAT-9 are percentile scores that compare each student’s ability to this nationally normalized sample. Therefore, if a student receives a score of 65 on SAT-9 math, the ability they demonstrated on this test supposedly ranks them above 65% of their peers in the normative group. It does not necessarily compare them to everyone else taking the test! However, Rogosa has found that these scores can be quite inaccurate. For example, Rogosa gives the probability that a student who really belongs at the 50th percentile on the 9th grade SAT-9 math test being more than five percentile points away from this score at 70%. That is, the chances that a student will be within five points of a true score are only 30%! While the percentage does improve as the tolerance increases from 5 to 10 or 15 percentile points and as the student’s true score changes, since this test normalizes scores around the 50th percentile most performance evaluations fall near this score. Rogosa reports similar results were a student to test and then retest on the SAT-9 math test. The probability that a student who truly belongs at the 50th percentile would, on a retest, score within 10 percentile points of his or her first test is less than half. (Rogosa 1999)

Accuracy of the percentile rankings aside, use of the SAT-9 in California is problematic for another reason. While the SAT-9 is normalized to a population in which 2% of the test takers are Limited English Proficient (LEP), in California almost 25% of the population taking the test are not yet proficient in English. Given the statistical assumptions of population similarity necessary to compare the two populations, such large discrepancies make comparisons between California and the national sample suspect. (Popham 2000) Nevertheless, California’s Superintendent of Public Instruction, Delaine Eastin, noted that all English learners took the SAT 9 test in English, and, as expected, they scored significantly lower than students who are proficient in English.
"The good news is that results for our English learners," she stated, "although lower than results for English proficient students, increased in almost all subjects and grade levels." (California State Department Of Education 2000) However, what inferences are to be drawn from this "good news?" Without additional information, the superintendent would have a hard time knowing why these student's scores "increased." Have the LEP students learned English? Are they learning the test? Are they learning the content tested by the SAT-9?

The second similarity between the SAT-9 and NAEP is, surprisingly, the lack of conformity between the content tested by the SAT-9 and the content specified by the State of California (or the other states) as important in state standards and frameworks. As previously mentioned, the SAT-9 is a commercially available test, produced by the publishers of other educational materials. It is used by numerous states in the U.S. As such, this test must cover a broad curriculum and cannot focus on the curricular content mandated by a specific state. Inevitably, the SAT-9 will either test students on material they have not seen, or will be overly focused on specific parts of the curriculum. Barton finds that most states use assessments that are not aligned with their educational standards. (Barton 1999) It should be noted that California has published its intention to develop its own state test in the future, but has decided to continue to use the results of the SAT-9 to provide stakeholders data on which to base inferences and to make policy until that future test is developed. This decision has other ramifications.

It is widely known that teachers will change tests to match curriculum, or, if that is not possible, will change curriculum to match tests. (Dietel, Herman et al. 1991; Heubert and Hauser 1999) Resnick and Resnick have found this results in a narrowing of the curriculum and over emphasis on "basic skills." (Resnick and Resnick 1992) This is especially true when the tests are high-stakes like the SAT-9. In fact, a National science Foundation study showed that teachers in classrooms consisting of high percentages of underrepresented minority students were significantly more likely to state that standardized tests caused them devote far more time teaching to the test. The National Academy of science concludes that this does not improve the overall level of education. (Harvard University 1999) In fact, The National Council on Testing and Public Policy (NCTPP) suggests that direct teaching to these tests exists in virtually all the states and, therefore, test results are inflated. While the test is kept secret to discourage teaching to the limited scope of the test, inevitably students and teachers "learn" the content of the test.

A perpetual problem with all the high stakes tests used by the states is that, over time, the limited content domain of the test is discovered. Once this domain has been exposed, and is taught to the students, test scores inevitably rise. In Figure 2-5, Linn recounts the work of Koretz, Linn, Dunbar and Shepard (Linn 2000) showing Grade 3 standardized math test results over a five year period. After the district changed test vehicles, scores initially dropped. However, over the ensuing years, scores again returned to their previous high level. Linn argues that "standardized test results in high-stakes accountability systems were yielding inflated impressions of student achievement." (Linn 2000)
Figure 2-5. This figure represents five years of median grade equivalent scores from a 3rd grade standardized math test. Norm Referenced Test 1 (NRT 1) was given in 1986 and 1990, Norm Referenced Test 2 (NRT 2) was given in years 1987 – 1990. The typical “saw-tooth” pattern emerges as students and teachers learn the test. When an old test (NRT 1 in this case) or a new test (ALT 2) is again introduced, student performance tends to more closely align to the results from the first year of the current test than with the results from the present year of that test. (Source: (Linn 2000), p. 8)

The trend in Figure 2-5 is apparently not atypical and Linn, in fact, suggests it is well known by policy makers.

Poor results in the beginning are desirable for policymakers who want to show they have had an effect. Based on past experience, policy makers can reasonably expect increases in scores in the first few years of a program...with or without real improvement in the broader achievement constructs that tests and assessments are intended to measure. The resulting overly rosy picture that is painted by short-term gains observed in most new testing programs gives the impression of improvement right on schedule for the next election. (Linn 2000)

While the repeatability of such a trend suggests it occurs regardless of population cohort, the fact that the set of students changes from year to year should not be ignored. It is quite common, in fact, for policy makers and the media to report improvements in test scores from year to year by comparing, for example, last year’s fourth grade to this year’s fourth grade. The two populations are not necessarily statistically similar and should not be compared. Rather, it is usually more appropriate to compare last year’s fourth graders to this year’s fifth graders by matching the performance of each individual this year with his or her performance last year.

Although one might expect that, given their implicit constitutional mandate to oversee education in the United States, the states would be in the best position to validly evaluate student learning and understanding, current practice suggests otherwise. In most states, and in California in particular, assessments of student achievement are not aligned with curricular standards and so do not necessarily test what is being taught in the classrooms.
of the state. In these situations, publishers and test designers, not policy makers or educators, determine what should be tested and, ultimately, what gets taught. Given the high stakes nature of these tests, research suggests tests drive curriculum. Furthermore, given the low accuracy of SAT-9 scores, their inevitable increase from year to year as the test is used, and the fact that, at least in California, the population the test is normalized to does not represent the population of students being tested, the inferences made about student learning and understanding from the data produced by the SAT-9 seem untenable. Finally, the inferences made about student learning by comparing SAT-9 scores between two populations, especially given the differences between the normalized and state populations, should be especially worrisome. That the state continues to justify the validity of its inferences about the learning of Limited English Proficient students using a test entirely in English suggests, at least for this quarter of the population, stakeholders have little hope of truly knowing what these students have learned in the state supported system of education.

2.7.5 Grades

Heretofore, I have cited literature suggesting that no test is a direct measure of innate aptitude or ability and inferences based on the data from various forms of high and low stakes testing alone are inherently flawed. Primarily, this resulted from a lack of content validity. Nevertheless, most of the tests reviewed consist largely of multiple choice items or are one time events. By their very nature, tests, especially multiple choice tests, sample a small portion of what someone is able to recall at the time they take the test. This is especially true when the ability of a student is judged based on a single test score. “More than 50 years of studies illustrate that neither scores on well developed tests nor any other single indicator can predict how school or workplace performance will differ from one person to another. This is because other important factors (seeking the help of others, work habits, etc.) are not reflected in test scores.” (AAUW 1992) Wiggins and others argue that, because of such shortcomings, assessments of student achievement and ability should be made on a local level, by a student’s teacher. (Wiggins 1993)

However, the literature on the accuracy of local grades is, in a word, contradictory. Shepard cites Dolan (1994) as suggesting that little research has been done assessing the validity of teacher grades, but that high school GPA correlates well with college freshman grades (r = .48) (Shepard 1993) and Linn found a similar correlation between these same two variables (r = .5). (Linn 1982) Nevertheless, how teachers arrive at these grades seems to vary widely, and what constitutes excellence in one classroom may not be the same in others. Popham suggests that the research concludes teachers use a “hodgepodge” of grading systems based on assessments, student effort, student attitude, conduct, and academic growth to arrive at final grades. (Popham 2000). However Haury cites Barron as concluding, “even though teachers know that the students who are most knowledgeable...are not necessarily the ones that get the highest grades, most continue to depend on multiple-choice test scores to determine grades.” (Haury Jul 1993) Wiggins claims that teachers have been forced to use such tests because of the external imposition of similar high stakes accountability tests. Such a correlation was also documented above. Furthermore he finds that teacher-designed tests end up becoming the “sloppy and arbitrary second cousins of the ‘secret’ standardized tests,” and that the grades teachers
give become increasingly less justifiable. (Wiggins 1990) The College Board argues that, "a consistent pattern of ethnic and racial disparities has been found across a variety of...educational measures used for high stakes decisions, such as high school grades, class rank, and indices of school quality and rigor of courses completed, as well as educational outcomes (e.g., college grades, persistence, graduation)." (College Board 2000) In fact, the research suggests that while female students tend to have better grades than male students, males tend to score better on multiple choice tests. (AAUW 1992) Surveys have shown that teachers use tests they make themselves, or the tests that accompany the instructional material provided by private publishers. (Barton 1999) While one might logically conclude that tests provided with specific curricular materials would have a high degree of validity, and especially content validity in particular, Messick cites DeCorte, et al. who conclude the use of tests supplied by textbook publishers seem to provide little in the way of validity. (Messick 1989)

The way data is generated is not the only important aspect one must consider in an evaluation of the accuracy of grades. The weighting of various measures to produce the final grade also needs to be considered. Howe referred to this as across criterion validity. Messick claims the scoring model used most widely by teachers is the accumulation of scores and warns that it should not be taken for granted that such a model is always appropriate. (Messick 1989) One must consider the goal of the assessment. Loevinger argues that cumulative models differentiate individuals by degree (a more normative model) whereas a classification model differentiates individuals by kind (a criterion-like model). In the former, the higher the score the more of an underlying construct one is said to possess. In the latter, the higher the score, the greater the probability one belongs in a certain category. (Loevinger 1957) As indicated above, which of these models an educator chooses is probably heavily influenced by external factors such as state teaching frameworks and standards, or an educator’s personal belief in his or her grading system.

Krathwohl, in the second volume of the Taxonomy of Educational Objectives, suggests that ultimately there is an erosion of the original intent of course evaluation so that what can be explicitly evaluated and that which can be easily taught verbally gets assessed. Furthermore, he suggests that students learn quickly which responses will be rewarded and which will be penalized. (Krathwohl, Bloom et al. 1964) Wiggins argues, along these lines, that the problem is not the use of grades, but that there is a single grade, with no clear, agreed-on, and stable meaning to summarize all aspects of complex performance. (Wiggins 1996)

As was the case with all other ubiquitous measures of student learning and understanding we have considered, the validity of inferences made on the basis of grades appears questionable. However, validating the inferences made from the assessments created by individual teachers would be difficult and costly. These assessments are usually limited to small populations of students and are specific to a particular curriculum. In addition, because teachers return most student work as feedback to the students, they usually only maintain a record of aggregate measures, not the actual work or specific sub-scores underlying that grade. The lack of generalizability from teacher to teacher is the argument many have used to justify the need for more standardized, large-scale tests from the time of Mann. However, as has been shown, the research suggests imposing these tests
narrow the curriculum, especially for underrepresented minority students, to preparing students to take a test. Some have suggested that student rank, teacher mean grade and teacher grade variance might offer an alternative to high stakes tests and a way to standardize the meaning of teacher assessments. (Crouse and Trusheim 1988) Others have proposed alternative forms of assessment.

2.7.6 Alternative Assessments

The National Center for Fair & Open Testing claims there are two fundamental problems with student learning in the United States that high stakes testing does not address. First, “U.S. students in general do not do well at analysis, synthesis, evaluation, application of knowledge or problem-solving – that is real thinking in subject areas.” Second, that students, disproportionately from low-income families and children of color, receive a really inadequate education...Their schooling is already dominated by teaching to “basic skills’ tests.” This organization claims that “assessment based on student performance on real learning tasks is more useful and accurate for measuring achievement – and provides more information – than multiple choice tests.” (FairTest 2000) Archbald and Newmann argue that the best way to discover how students think, or to diagnose where they are having difficulties in learning – which, aside from accountability and placement, is the main reason for testing – is to give them as much range as possible to express themselves fully. (Ascher 1990) The literature agreeing with this basic proposition is extensive. For example, Marc Tucker suggests that the fundamental unit of an examination system ought to be the task. “You ask the kids to perform tasks that might take them [some time] to address,” and construct an examination system out of tasks that are as close as possible to the kinds of behaviors that you would want children to be able to exhibit as adults. (Tucker, Sizer et al. 1992) This is quite different than standardized testing in which, according to Bowers, the main purpose is to sort large numbers of students in as efficient a manner as possible. (Bowers 1989) Marks suggests that students are more engaged in learning in classrooms that use authentic assessment (Newmann, Secada et al. 1995), and Wolf suggests that these tasks can actually become “episodes of learning” for students. (Wolf 1993) Weiss and Weiss suggest that even principals and teachers are “becoming frustrated with conventional evaluation practices typically used to determine teacher effectiveness...such as lecture, demonstration, recitation, and modeling designed primarily to transmit knowledge and cognitive skills to students.” (Weiss and Weiss 1998) The NCTPP argues that alternative forms of assessment must be developed to promote the greater development of all. They continue that too many tests, emphasizing lower order thinking skills, are devouring valuable instruction and learning time. “We need to move to tests that can be used by individuals and institutions to promote human development.” (NCTPP 1990) Stiggins observes that with enhanced visions of the complex nature of the meaning of academic success came a sense of the insufficiency of traditional multiple choice tests. (Stiggins 1995) Neill notes that the greatest weakness the TIMSS identified about American students was not that they hadn’t mastered the basics, but that they didn’t understand concepts and how to apply knowledge. (Neill 1998) But how we go about implementing such a process is an open question. While Haury agrees that we must develop assessment tools that allow students to demonstrate skills and perform observable acts, he notes no agreed upon way to introduce such tools.
He states that while there is an urgency "to develop innovative forms of assessment to provide valid indicators of student understanding and other learning outcomes," there are diverse opinions on how to proceed. (Haury Jul 1993) Perhaps an accepted definition of what alternative assessment encompass might be a good place to start.

The Congressional Office of Technology Assessment (OTA) defined alternative assessment as "testing methods that require students to create an answer or product that demonstrates their knowledge and skills." (OTA 1992) Federal law has further codified the definition. The Goals 2000: Educate America Act states that "performance standards" should embody concrete examples and explicit definitions of what students have to know and be able to do to demonstrate that such students are proficient in the skills and knowledge framed by content standards; (Goals 2000: Educate America Act. Sec. 3(a)(9)) Herman, Aschbacher, and Winters emphasize that "these new assessments stress the importance of examining the processes as well as the products of learning. They encourage us to move beyond the 'one right answer' mentality and to challenge students to explore the possibilities inherent in open-ended, complex problems, and to draw their own inferences." (Herman, Aschbacher et al. 1992) Yale Professor Emeritus Edmund Gordon suggests that, "the task then is to find assessment probes (test items) which measure the same criterion from contexts and perspectives which reflect the life space and values of the learner." (Dietel, Herman et al. 1991) Herein seems to lie the rub – can one make tasks that are equal and equitable?

Lam states that "there is little research devoted to examining and promoting fairness in performance assessment." He illustrates the dichotomous nature of the problem of fairness in testing by pointing out that equality in testing (i.e. standardization) ignores the fact that bias will exist if different groups have irrelevant knowledge or skills that can affect assessment performance. This is Messick's construct irrelevant easiness discussed earlier (Messick 1989). On the other hand, having equity in testing by testing tailored to the individual student's instruction, background, etc., runs the risk of producing results that cannot be compared. Dietel, et al. cite Linn as observing that "it is a mistake to assume that shifting from standardized tests to performance-based assessments will eliminate concerns about biases against racial/ethnic minorities or that such a shift will necessarily lead to equality of performance." (Dietel, Herman et al. 1991) In addition, Sanders observes that there is great difficulty generalizing from a performance that is highly task dependent to the larger realm of the subject area being assessed, and argues that this difficulty is, in fact, consistent with cognitive research suggesting thinking is situation and context-specific. (Sanders and Horn) Barton observes that while multiple choice tests cannot assess a student's ability to come up with their own answers, the performance assessment movement has been slowed over problems of reliability and measurement error. (Barton 1999)

There are other downsides to these alternative forms of assessment that must also be acknowledged. According to Quellmalz, "Performance assessments are generally valued for testing students' deep understanding of concepts and inquiry strategies, for making students' thinking visible, and for measuring skills in communicating about their...knowledge. In addition, performance assessments can present authentic, real-
world problems that help students to show they can apply academic knowledge to practical situations. On the other hand, performance assessments are time-consuming and costly to develop, logistically demanding, and of questionable utility if not developed and scored according to sound measurement methods.” (Quellmalz, Schank et al. 1999) In addition, others have wondered if these types of assessments might not lend themselves to inappropriate coaching just as current testing formats do. In fact, Madaus relates his experience in Ireland where students were prepared for essay tests by memorizing an essay that could be rewritten in slightly different ways, depending on the topic called for. (Ascher 1990)

As with the large scale tests and grades previously reviewed, performance based assessment formats are not without their shortcomings. They tend to be expensive to implement in terms of time and money, can be hard to distribute, and, as was discussed in Section 2.4.1 The Third International math and science Study, can also be culturally and contextually inappropriate. Finally, it seems that the validity of the data from these types of assessments has not, for the most part, been established. The literature, rightfully so, warns that just because the performance of tasks may allow students to demonstrate and apply what they really know, if the tasks are unable to generate the data necessary to support the inferences made, those inferences will be invalid. Additionally, depending on the tasks developed and used, racial, gender and ethnic biases may still be apparent.

2.8 Conclusion

Our inability to accurately assess what students learn and understand adversely affects national and state educational policies. It lessens the learning of students. Currently more than 50 individual standards and frameworks exist in the US educational system. As a result textbook publishers must include material that addresses these individual standards or at least the standards of the states and territories where the majority of texts are sold. The resulting curriculum has been termed a “mile wide and an inch deep.” As the states remain adamant in their demand to oversee education within their borders and the U.S. constitution does not specify education as a federal power, this situation appears immutable in the near term. Tests that propose to evaluate student achievement and learning on the national level cannot hope to validly cover the content individual students have been exposed to in such a cacophony of various curricula. Therefore, as cited above, we have little credible information about what “works” and what does not nor are we able to determine what in the present system of education is “effective.” As a result, most states are developing a unique high stakes competency test for its citizenry. Usually these tests are intended to perform gate-keeping functions such as determining whether a student is allowed to move on to the next grade or to graduate from high school. Due to the importance of these tests, teachers report spending a lot of time preparing their students to take such tests. In effect the assessment drives the curriculum. While these tests allow policy makers a “bright line” cutoff to determine whether to promote a student to the next grade, they suffer from many of the same problems cited for the national tests such as classification error, relying on a single score to determine important outcomes, and content invalidity. In addition, these tests cannot be used to compare students or schools across state boundaries. As a consequence, federal policy-makers rely on
anecdotal experiences based largely, it seems, on media reports, low-stakes sampling like the NAEP or single high profile events such as the TIMSS to inform their decisions. State policy makers are at a similar loss to compare their systems to those of other states or nations. Even teachers have little valid evidence with which to objectively compare their students with the students of other teachers or to the state standards.

The objective of this chapter was to review the pertinent literature on testing in America and to suggest why we have had such problems accurately measuring what our students learn and how our teachers teach. More specifically, it has attempted to briefly explain the development of our current testing system, and the limitations on what we can conclude from those tests. The chapter highlighted how various validity measures developed over the last century to help avoid misinterpretation of data and inappropriate decisions based on false interpretations. Unfortunately, the bodies that have, for the most part, developed these “standards” have no legal authority to monitor or enforce compliance.

The limitations of the current system of assessment in the United States deny educators, policy-makers and researchers the ability to accurately evaluate student learning, assess the educational systems they are responsible for, or describe the system they want to study. As a result, education policy in the United States has oscillated between “new” reforms and return to “basics” for the last half century. There is little in the way of evidence to confirm that either of these policies have had any positive or negative effect on how well teachers teach and student learn or understand the subjects they study. More importantly, we have little solid evidence to suggest ways to improve the system for future generations of Americans. This is not meant to imply that every testing system explored above is invalid. As was made clear, tests are neither valid nor invalid. It is only the inferences that are made from test data that can be so classified.

Excepting teacher grades and performance based assessments, the inferences made from the data generated by the instruments discussed have been the subject of extensive validity and reliability studies. Used to support the inferences they were designed to inform, all these tests can tell us something, even if only how well U.S. students recall the basic math facts an international committee feels are critical to recall. However, these vehicles generate quantitative data that seems just too good not to use in inappropriate ways or in ways that have not been validated. For example, it is easy to compare two scores, even if those scores represent groups or individuals that are statistically disparate, have not been exposed to the same curriculum, have different abilities to read the questions posed, or are likely to do better or worse because of the format of the questions. Nevertheless, the research strongly suggests that we might be able to use data from all these sources to “triangulate” on what a student really knows and understands, and then use those inferences to truly improve how teachers teach and students learn. What is needed is a tool to assess authentic student learning holistically and that can generate data from which policy-makers, teachers and students can draw accurate inferences. Ideally, the data generated by this tool should correlate to some degree with data generated by present assessment methods.
Large-scale assessments are designed to cover a wide range of content relatively quickly and can sort large numbers of students very effectively. When appropriately tied to curricular content, they seem to allow quick evaluation of the facts a student knows. Generally, these tests allow one to predict future performance to varying degrees, they are cheap to administer, and scoring is usually quite objective, even though what actually is on the test is more subjective. It is generally quite easy to rank students with these types of tests and they are often designed with ranking against a normative group in mind. Most importantly, they are ubiquitous and the demise of such testing systems does not seem imminent as the results they report seem to be widely accepted, even demanded, by the public and policymakers.

Nevertheless, the use of these tests alone to generate the data upon which US policymakers, educators, parents and students make decisions about the education of millions of individual students has not, it seems, served us well. Over the last fifty years, the nation has moved from one fad to another as it has responded to the latest “crisis”. The literature not only suggests new assessments of student learning and understanding are warranted, but also suggests what that system might be – performance assessments. (Schloemer 1997) George Madaus, director of the Center for the Study of Testing, Evaluation, and Educational Policy, believes that performance-based testing "is not efficient; it's expensive; it doesn't lend itself to mass testing with quick turnaround time – but it's the way to go" (Bowers 1989)

Wiggins, and a number of others, have argued that we need to allow students to express themselves fully in order to really get at what they know and understand. (Gardner 1983; Ascher 1990; Wiggins 1993) Jorgensen claims that “performance-based assessment requires that the students complete, demonstrate, or perform the actual behavior of interest. There is a minimal degree of inference involved.” (Haury Jul 1993) Archbald and Newmann have suggested that the "problem solving model," can be adapted to almost any knowledge-based discipline. It involves the presentation of a problematic scenario that can be resolved only through the application of certain major principles, theories, or formulae that are central to the discipline under examination. (Bowers 1989) Nevertheless, to implement a system of performance based assessment one must overcome some relatively challenging problems.

In her paper, *Connecting Performance Assessment to Instruction*, Fuchs (Fuchs 1995) argues that these are the minimum requirements for performance assessments: That they

1. Measure learning outcomes that are based on real-world performances and that are critical to the course of study.

2. Address the need for formative, summative and placement decisions

3. Provide clear descriptions of student performance such as problem solving strategies, and student competencies that can be linked to instructional actions and generate profiles of student competence.
4. Be compatible with a variety of instructional models.

5. Be easily administered, scored, and interpreted by teachers. Performance assessment must be manageable in the school setting.

6. Make the goals of learning clear to teachers and students. In particular the scoring of student performances should be clear, concrete, and visible.

7. Generate reliable, valid, and meaningful information.

Winograd also argues forcibly that we should not design systems that impose a specific structure on student solutions because such systems are bound to fail as they exclude valid solution representations or force students to mimic representations that do not accurately represent their understanding. (Winograd and Flores 1986) To avoid this, the assessment should model student understanding in a manner that allows the student to build a solution based on his or her own representation of how to solve the problem. Such an assessment would allow us to more fully examine what Norman refers to as the “mental model” of the student. (Norman 1993) Therefore, such an assessment must allow for multiple representations of a solution. In doing this, we approach assessment from a fundamentally different direction than the test design paradigms currently in use, especially those used to develop multiple choice type tests (see for example (Mislely, Steinberg et al. 1999)). Nevertheless, this approach is similar to current methodologies in its need to have an understanding of the content and criteria against which students will ultimately be assessed.

Finally, as Schwartz argues, if these assessments are to approximate “real-world” activities they must be public. (Schwartz 1990) In effective workplaces, what is expected of you and how you will be required to meet those expectations are generally well known. Under current educational assessment systems however, what will be assessed and how a student will demonstrate his or her mastery of a concept is, most often, kept secret until the moment of the test. This secrecy is largely for the convenience of the test maker not the test taker, since only small samples of the domain will be tested. (Wiggins 1990; Wiggins 1993) As has been demonstrated however, this does not ultimately prevent the practice of teaching to the test. On the other hand, if a very large number of authentic tasks were developed, and the instructional domain was adequately covered, teaching to the test would be teaching the curriculum. Assessment would no longer be as much about trying to guess what would be included in or excluded from an assessment, rather it would be building proficiency in learning and understanding a known curriculum. In addition, because students could be subject to assessment on any particular item, the goal would become developing proficiency at successfully accomplishing all similar tasks. Tests and test items could be integrated into the curriculum instead of being distinctly separate from it. Furthermore, test items would be available for public inspection and comment prior to practice or testing. In this way, biases or inaccuracies within problems could be identified and corrected before being used.
As Fuchs and others have suggested, the grading rubrics for these tasks should also be public and understandable. (Wiggins 1993; Fuchs 1995) Resnick and Resnick suggest that if the results of NAEP writing assessments are any example, reliable, publicly believable, quantitative measures can be derived from judgments of these assignments. (Ascher 1990)

The interconnectivity, malleability, information storage, and analysis tools afforded by modern computer technology are well suited to address many of these requirements. As such, promising new sources of data, which describe actual student learning and understanding, are arguably available to improve our inferences of true student achievement.

I now turn to how modern computer technology not only allows us to develop and deliver these performance-based assessments easily and cheaply, but, more importantly, how the data such assessments provide can dramatically improve our inferences of student understanding and learning.
Chapter 3. Assessing student learning and understanding with technology

As discussed in the last chapter, the literature suggests some specific needs one must address in order to improve the inferences made from the data generated by tests and assessments. We must be attentive that a test or assessment covers the breadth and depth of what was actually taught. We must not overly restrict a students' expression of how they construct an answer if we are truly interested in understanding how they think about a concept. We must use an assessment tool that actually predicts some near or long term student performance that we are interested in. In general, the field has classified these needs into three categories: content, criterion and construct validity. Moreover, the literature also suggests that in order to really evaluate what a student knows and understands, we need to observe how students actually go about solving authentic problems without imposing a pre-determined assessment model on such a solution.

Finally, both the problem and evaluation rubrics need to be public before, during, and after assessment. The literature suggests these fundamentals must be addressed to assure accurate inferences about student learning and understanding, regardless of the form of the test or the amount of technology used to assess students.

The adaptive nature of modern computer technology, its ability to store large amounts of data, and its capacity to quickly and accurately process information suggest modern computer technology could have an important place in educational assessment. It seems that, if employed properly, this technology could be an effective tool in helping us satisfy the testing and assessment needs outlined above and in the previous chapter. Nevertheless, the promise and capabilities of this technology seem, at present, largely underutilized. In fact, because the technology has often been used in the same ways as pencil and paper, its application has often suffered from many of the same problems as tests outlined previously.

Using technology to improve assessment in our schools might not be appropriate if it is just used to do the same thing we are currently doing with pencil and paper. If computer software is used to record and score standardized, multiple choice tests, one might be able to rank students more quickly, but the content of the test may still not match the content of the instructional curriculum, etc. This is especially true with prepackaged software that cannot be adapted to a course of instruction or software that repeatedly drills students until they have memorized enough information to perform well on a particular multiple-choice test. Moreover, we might be measuring some “innate ability” but that measure might tell us very little (or nothing at all) about how a student solves the problem we pose. If we don’t pay particular attention to what we assess, a student’s performance on one problem or test probably will not let us conclude anything about her or his performance on a similar problem or test in the near or long term future. In other words, we may be able to make invalid inferences more quickly and at much greater cost. These applications, however, do not fully exploit the power of computer technology, and do not improve our ability to assess student understanding.
In order to fully capitalize on the power of technology, the literature suggests software to assess student learning must be designed in new ways. For example, Sivin-Kachala and Biallo report that programs which provide multiple pathways to information are more valuable than those that impose single presentation standards, and they cite Bastecki and Berry as concluding that learners prefer to impose their own spatial organizing schemes on the information provided. (Sivin-Kachala and Biallo 1998) Research also suggests that technology be employed as a tool rather than as an evaluator. (Winograd and Flores 1986) For example, Radlinski and Atwood report that since intelligent tutoring systems don’t understand the context a student is working in, students did not like the tutor because it gave them advice that was inappropriate for achieving their goals in the way they wanted to. (Radlinski and Atwood 1998) Schwartz also reaches similar conclusions and suggests that computers should not evaluate student expression, but should merely respond to student input, showing what they have requested and letting the human make sense of it. In fact, he suggests that if the result does not conform to what is expected, cognitive dissonance is created in the student which encourages changes in her or his mental model. (Schwartz 1989) Along the same line, Wood found that when software was used as a tutor, a student’s algorithmic application improved, but when students used software as a tool, or an Intellectual Mirror as Schwartz calls it, their understanding improved. These findings suggest that the technology we need in order to gain more accurate insights into student learning and understanding acts more like a passive tool in the hands of a student rather than a substitute for human evaluation. An educator could use such a tool much like a physician uses X-ray or an MRI. These tools inform the physician; they do not “make a decision” for her or him.

In order to test whether technology can improve the accuracy of our inferences and provide new insights into student learning and understanding, a different kind of software must be found that exploits some of the more powerful capabilities of computer hardware and software than is currently the norm. The software should allow teachers to adapt the content of the assessment to his or her curriculum in order to assure agreement between what is taught and what is tested. The software should allow teachers to test what a student knows and assess how well the student can apply that knowledge to the task or problem at hand. The tasks or problems might not only be those that meet necessary standards and frameworks, but also the authentic tasks and problems that students can realistically be expected to encounter in their lives. The inferences we make about a student’s ability to accomplish these tasks should be predictive of future performance on similar problems. Finally, the software should allow us to see not only a student’s answer, but also how s/he arrived at that answer so we can begin to make inferences about what the student truly understands. Obviously, the software tool should not require the student to solve a problem in a particular manner or evaluate student performance against some pre-defined algorithm, rather it should merely report how the student went about solving a problem and what kind of success a student had at solving the problem.

3.1 The Interactive Multi-Media Exercises (IMMEX™) software

Traditional forms of pencil and paper tests often lack the ability to accurately record how a student solves a problem from its presentation to its solution. While this is especially
true for multiple choice type tests, the transitions students make in solving open-ended problems can also be unclear. Some early complaints of computer tutoring or learning systems (CTS/CLS) emphasized the need to go beyond using computer technology as just another multiple choice assessment system. Many designers have attempted to improve these systems by giving them "intelligence" in the form of templates that "experts" would use to solve the problem. Primary criticisms of these "expert" systems include their inability to measure novice problem solving (Nickerson 1995), the difficulty discerning exactly how experts actually solve a particular problem (Bloom, Wolff et al. 1998), and the limitations these systems place on what is considered an "acceptable" student solution strategy. (Lesh and Kelly 1996) Another criticism of these systems is that they are often designed by software engineers unfamiliar with or having limited experience in education or pedagogical theory. The Interactive Multi-Media Exercise (IMMEX™) software overcomes these and other shortcomings.

The IMMEX™ software was developed by Professor Ron Stevens as a problem presentation and student assessment tool for use in his Immunology classes at the UCLA medical school. Palacio-Cayetano (Palacio-Cayetano 1997) reports that, among others, the primary goals of IMMEX™ are to:

1. Increase students’ ability to problem-solve by using higher order thinking skills.
2. Enhance students’ ability to integrate relevant information from multiple sources while problem solving.
3. Improve the ability of teachers to evaluate students’ problem solving skills.
4. Enable teachers to customize their curriculum by using new technology in a meaningful way.

To fulfill these goals, the IMMEX™ software consists of three modules: an authoring module, a presentation module, and an assessment module.

3.1.1 The IMMEX™ authoring module

The IMMEX™ authoring module allows problem creators (usually teachers) to develop new problems for classes or subjects that are pertinent to the curriculum they teach or are interested in. The authoring module also allows existing problems to be modified so that they can adapt to meet the needs of different groups of students, teaching styles or instructional contexts. This malleability alone makes the IMMEX™ software unique. Individual teachers, teams of teachers, or other authors create IMMEX™ problems by specifying authentic challenges they want students to solve by allowing students to gather information (content) and apply what they know (strategies) to reach a conclusion. When creating problems, the authors first specify a problem that requires solving. This is called the prologue. Current IMMEX™ problems ask students to find the biological parents of a baby, identify the cause of a physiological reaction, or identify historical times and places, among hundreds of others. Problems have even been created in other languages both for English language learners and English speakers studying other languages. In addition to specifying the problem and the problem’s possible solutions, creators of IMMEX™ problems include the results of various tests on or conjectures about an unknown in the problem, and reference information that may or may not be germane to
the solution of the posed problem. The sum total of all information in an IMMEX™ problem represents the “problem space.” Hurst, et al. indicate that “conscious efforts are made during the problem design to make the problem space as broad as possible to encompass a multitude of student [solution] strategies.” (Hurst, Casillas et al. 1997) In other words, the problem space is designed to be flexible enough to allow students the opportunity to approach problem solving in a way that makes sense to the individual student.

Authors can also assign “costs” to each test or conjecture a student selects. These “costs” represent the time or money such tests would actually cost the student if they were ordered in reality. In this way students are faced with the added and realistic constraint on resources that exist in real-world problem-solving. The student’s final score reflects how well they used available resources. This score feature was not used in this study and will not be addressed further.

Finally, the author(s) of an IMMEX™ problem can add an epilogue at the end of the problem. When present, the epilogue usually poses critical questions that had to be answered in order to arrive at the solution to the problem and where that information could have been found in the problem space. While it is possible for an author to specify a single algorithmic solution to the problem in the epilogue, none of the IMMEX™ problems I have reviewed for or used in this research detail such algorithmic solutions.

The design of the IMMEX™ authoring module allows problem authors to easily create numerous versions or “clones” of an IMMEX™ problem by making small changes to the prologue and to various pieces of information in the problem space. For example, in the IMMEX™ problem My true biological roots a student is asked to help a classmate determine if the parents she was raised by are really her biological parents. By ordering various genetic tests, students can identify the true biological parents of the classmate. After the authors designed the first version of this problem, they made an identical copy and subsequently rearranged the results of certain tests to point to a different solution. In this case, the information in clone one confirmed that the parents of the classmate were her true biological parents, while the information in clone two confirmed another couple to be the classmate’s true biological parents. By cloning problems in this manner, problem authors can easily develop dozens of similar clones from a single initial problem.

3.1.2 The IMMEX™ presentation module
The IMMEX™ presentation module delivers an IMMEX™ problem as a series of web pages via the world-wide web. As such, IMMEX™ problems are platform independent. The prologue is always the first page delivered. It describes the problem context for a student, what problem or problems the student is expected to solve and any other pertinent information the problem’s author thinks it desirable or necessary to provide. For example, sometime “hints” as to how to begin a problem may be included. The student then selects a path through the problem space s/he feels is best to reach a solution to the problem. In IMMEX™, the path to a solution is comprised of a series of steps. Each step
involves a student choosing one of a number of menu items from pull down menus or hyperlinks in a standard graphical user interface or web page. These items of information usually represent the results of a test conducted or conjecture made to determine information about an unknown. These "tests," for example, could involve blood tests to determine genetic information, the time it takes for a sonar signal to travel between two points, or a line of dialogue from a play, depending on the problem being solved. The key point is that the student "orders" a test and receives information, which s/he must then interpret in order to solve the problem. (See Appendix 9) While a number of tests may yield important information about the problem's solution, other tests will yield duplicate, minimal, or no important information. Therefore, while multiple paths to a solution exist, it is entirely up to the student to transit the problem space in a way that makes sense to him or her. The path from problem start to solution is not predetermined; however, the problem does have some structure in that the problem author(s) generally group similar tests together under a specific menu item or hyperlink. In addition to informative tests, students may also choose to view reference materials as they are trying to solve the problem. These reference materials may include content information a student was expected to "know," (e.g. the flame color of burning potassium) or technical information most problem-solvers would have to look up (e.g. the refractive index of silica glass).

To solve the problem, a student can either select from a list of possible answers or type a solution into a text box. The answer mode is determined by the author of the problem. Also, while a student can have one or more chances at correctly solving the problem, most IMMEX™ authors set this parameter at two tries, as that allows the student to recover from mistakenly choosing the wrong answer, but prevents frivolous guessing. Here again, while the author of the problem determines this value initially, it can be modified by other authors depending on the context in which the problem will be used. Finally, because most IMMEX™ problems have multiple clones, these clones can be presented to students in any order desired. Generally, the order a particular student receives a clone is entirely random. Consequently, as a group of students work through a specific IMMEX™ problem, each student will be working on a different clone and will, therefore, receive information that supports a different solution than a peer simultaneously working nearby (either physically or in cyberspace). Nevertheless, the similarity between the problem spaces of distinct clones allows educators to use a single IMMEX™ problem longitudinally as both a practice and an assessment event. This similarity will be discussed in greater detail in Chapter 4.

3.1.3 The IMMEX™ assessment module
IMMEX™ represents student problem-solving performances either graphically or numerically. The graphical representation presents each item of information as a distinct, labeled rectangle. Each information item in the problem space is represented by a unique rectangle. The visual representation of the entire problem space for the IMMEX™ problem Hazmat Holiday Special (a qualitative chemistry problem) is shown in Figure 3-1 as an example. This representation is termed a problem space "template." Since rectangles representing similar types of information are usually grouped together under a specific menu item or hyperlink, they are generally given the same unique color and are
grouped together on the template. Such a grouping is intended to make interpretation of student solutions easier; however, an assessor may change these qualities to accommodate different assessment needs.

**Figure 3-1.** The graphical representation of all possible information items (tests and references) a student could "order" while solving the IMMEX™ problem called *Hazmat Holiday Special*. This graphical representation of the information universe is called a template. Although not fixed, rectangles are usually colored to represent specific types of information and are usually grouped together in the same way they were in the problem space presented to students. For instance, in this case rectangles along the left side of the figure represent reference items, while rectangles in the center of the picture represent physical tests that can be performed on an unknown compound. Rectangles on the right side of the template represent chemical tests that a student can conduct on the unknown material.

As described above, when a student solves an IMMEX™ problem s)he moves from one information item to the next. Concurrently, the assessment module records these steps and it builds a unique graphical representation of the student’s solution based on the general problem template. This unique representation is called a "search path map". Initially, the search path map consists only of a "start" box representing the problem prologue. As the student proceeds through the problem space toward a solution, IMMEX™ adds a new rectangle (representing each new piece of information a student viewed) to the graphic along with a line indicating the order in which the items of information were selected. As illustrated in Figure 3-2, a line extends from the upper left of the rectangle the student went from to the lower center of the rectangular item the student went to. The student performance always begins at a start box (in the upper left hand corner) and ends at the solution box (usually along the lower edge of the graphic). For example, in Figure 3-2, after reading the prologue, the student examined information item number 10 (the box at the upper right of the figure). Next the student investigated menu items 3 – 6, in order, and finally attempted to solve the problem (the black box at the bottom right of the figure).
Figure 3-2. This figure represents the series of steps a student used to solve a hypothetical IMMEX™ problem. Known as a “search path map”, the graphic shows what information or tests the student used to solve an IMMEX™ problem and in what order s/he “ordered” the information.

For reasons to be discussed in Chapter 5, computers can also be helpful in identifying various patterns in student problem-solving performances. To facilitate this type of processing, the IMMEX™ assessment module is also able to generate a numerical representation of a student performance. The numerical representation of the hypothetical student performance in Figure 3-2 is shown in Figure 3-3 below.

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>10</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 3-3. This figure gives the numerical representation of the hypothetical student performance represented graphically in Figure 3-2. In this case, the performance is for Student 001. The top row of numbers represents individual information items. “Start” (the prologue) is item 1 and item 11 (not shown) represents the student’s solution. The second row of numbers indicates the information items a student selected. Taken as an ordered pair, an item in the top row coupled with its corresponding item in the bottom row shows a single transition in the student’s problem solving process. For example, in this figure the student went from information item 1 (start) to item 10; then from item 10 to item 3 (the column on the rightmost side of this figure).

Figure 3-3 shows the performance of Student 001. This representation documents the progress of this student through the problem space in columnar form. After the initial presentation of the prologue, which is considered to be information item 1 here, the student decided to view information item 10. Therefore, the first column shows the student moving from item 1 to item 10. The student next moved from item 10 to item 3. This transition is represented in the far right column of Figure 3-3. After viewing item 3, the student moved on to item 4 (the third column of numbers) and so on, until s)he attempted to solve the problem. The solution attempt is represented as item 11 in the figure, and as shown in Figure 3-3, the student attempted to solve the problem after viewing information item 6. Whether the student actually solved the problem correctly is not shown in this representation.
Taken alone, the bottom row of numbers in Figure 3-3 represents a single performance vector or data point for Student 001 on this clone and numerically encodes how the student transited the problem space in the hypothetical IMMEX™ problem.

In addition to search path maps, the assessment module can also display a timeline representing the amount of time a student spent reviewing each item of information relative to the total time a student took to solve the clone. Here again, this capability was not used in this study.

3.1.4 Additional capabilities of IMMEX™

As currently implemented, IMMEX™ is delivered to users over the World Wide Web from a file server at the University of California, Los Angeles (UCLA) IMMEX™ lab (www.IMMEX™.ucla.edu). Although IMMEX™ can function on individual personal computers running the MS Windows® operating system, such a configuration increases the chance that data could be misplaced and it makes the processing of data more difficult since the data must be downloaded from multiple disks, reformatted, and integrated into the larger database. By delivering IMMEX™ problems over the World Wide Web, not only are all student performances recorded as a student makes the transition between information items in an IMMEX™ problem, but the student and the student’s teacher can also review student progress in real time via the World Wide Web. The IMMEX™ web site also provides a number of data analysis tools which authorized educators, policy-makers and researchers can use to assess student learning and understanding. I will address the tools relevant to this study in subsequent chapters.

In addition to student and teacher access, providing IMMEX™ problems from a central location via the World Wide Web allows any educator, policy-maker, or member of the public to review actual problems, but not individual student performances, at any time. As such, problems may be evaluated for content appropriateness, unintentional bias, or against other standards of applicability long before a problem or its clones are actually used in a classroom. Finally, when a problem is used, students with access to the World Wide Web from home, school, library or other entry portals can perform IMMEX™ problems at any time. The similarity of this system to that proposed by Schwartz is particularly noteworthy. (Schwartz 1990)

3.2 Study Methodology

The answer to the question of whether modern technology offers us the opportunity to gain more accurate insights into student learning and understanding depends, in part, on the methodology used to gather and analyze the data. It is the soundness of this methodology that will determine the reliability and generalizability of the conclusions drawn from the data. This section describes the methods used to collect the data underlying the conclusions described in subsequent chapters. It also details the demographics of the student population used in this study.
The methodology used in this research can be classified as "quasi-experimental." In an experiment, all extraneous variables are accounted for and controlled, and the independent variables of interest are isolated. As such, specific independent variables can be manipulated to discern a specific correlation between, or causal effect of, that variable and the dependent variable of interest. The evaluation of social or cultural systems, given the large number of independent variables, makes controlling for all variables beyond those of interest extremely difficult.

A quasi-experimental study is used in a situation in which it is not possible or desirable to control all variables. The later category includes situations in which it is important to allow for observation to take place in a normal context. Rogoff cites the study of learning as just such a situation because of the role student interaction and culture plays in learning. She also warns that observers must acknowledge the effect of their presence on the system when studying such systems. (Rogoff 1998) In quasi-experimentation then it is not necessary for the researcher to control all variables in a situation, but it is important that the researcher be attentive to the variables that can affect outcomes and to report the variables that appear to influence results.

Wolff, et al. (Wolff, Bloom et al. 1998) cite three different quasi-experimental methodologies researchers can use to conduct inquiries when variables aren’t otherwise controllable:

1. Compare non-equivalent groups.

2. Interrupted time-series, in which effects are inferred from comparing the measures of performance of one group taken many times before and after treatment.

3. Passive observations, in which researchers attempt to infer causation or correlation between two variables based on observations of events as they naturally occur.

A quasi-experimental, interrupted time-series format was chosen for this study for four reasons. First, as alluded to previously, field experimental studies in an actual classroom have various constraints, the chief of which is that the learning of the students not be effected in a negative way. Similarly, it becomes difficult to dictate the actual curriculum a professional educator should deliver, especially if the educator has taught for sometime and has developed a teaching style, scope and sequence with which s)he feels comfortable. Requiring that educator to change styles for research may control one variable, but may introduce more overall complexity such as teacher unfamiliarity with new methods, etc.

The second reason for choosing this methodology is that constant observation was not desirable. The educator I observed did not want a full time observer in her classroom, and it was felt the constant presence of a researcher would change the environment of the classroom. In addition, since the goal of the study was to address the accuracy of inferences and the development of new insights it was felt
that the classroom needed to be as similar to the prior semester as possible, with the exception of the addition of IMMEX™ technology.

Thirdly, it was neither desirable nor possible to control for many of the variables present in the typical American high school classroom. For example, assignments of students to a particular class, the time of day a class met, and how motivated a student was to succeed in the class were all variable beyond the ability of this researcher or even the teacher to control.

Finally, because the number of available students was less than 200, further dividing this group into sub-groups could yield cohorts so small that results would be difficult to substantiate statistically.

3.2.1 Pre-treatment
In a quasi-experimental, interrupted time-series study, establishing a baseline is critical. In this case, it was especially important to understand the accuracy of current assessments and the insights into student learning and understanding already possessed by the teacher, school, district and state.

3.2.1.1 Pre-treatment student survey and results
A student questionnaire (see Appendix 1) was administered to 134 students in four college preparatory high school chemistry classes at a large, suburban high school (Lincoln High School – a pseudonym) near Los Angeles, California. All students were instructed by the same teacher (Ms. Lindsay – also a pseudonym) for 50 minutes each day, five days each week. This was the second semester of the first year of high school chemistry for each student. The students answered the survey at the beginning of a regular class period on February 28, 2000. The student questionnaire was intended to determine the basic demographic make-up of the class and the students' feelings about certain variables that the literature suggests might effect assessment accuracy and the inferences that result from those tests. These variables included apparent motivation, realism of standards, relationship between student and teacher, memorizing for tests, etc. In addition, the questionnaire included a set of questions that probed student attitudes about various aspects of problem solving.

Of the 134 students responding to the survey, 57% were female and 43% male. Ethnically, 66% of the students considered themselves white, while the remainder identified themselves as Latino/Latina (10%), Asian/Asian American (10%), Multi-ethnic/Biracial (5%), or refused to state an ethnic group (8%). Given the small numbers of students in each of the ethnic groups other than white, it was impossible to investigate the statistical relationships between specific ethnic groups and other variables in this study. Therefore, for the remainder of this study, ethnicity will be considered a dichotomous variable and students will be classified as white or non-white.
Although, relatively ethnically homogenous, Lincoln High serves a slightly more diverse socioeconomic population as measured by median home price within zip code areas. Forty-one percent of the students live in a high SES zip code (median home price $445,000), 33% live in a lower SES zip code (median home price $310,000), and 18% live in an even lower SES zip code (median home price $283,000). In this area of Southern California, these home prices define the middle to upper-middle class. The other eleven students live in a number of slightly lower SES zip codes. Surprisingly, the distribution of ethnic groups among zip codes is random, although marginally so ($\chi^2 = 7.206, df = 3, p = .071$). The majority of students report that their parents are well educated, which is not surprising considering the median prices of the homes in which the students live. Seventeen percent of the parents have some college experience, but no post-secondary degree; 32% have bachelor degrees; 30% have masters degrees; and 5% of the students report their parents have earned a doctorate. Only 10% of the students reported that their parents did not continue their education after high school.

The students in the sample population range in age from almost 16 to slightly over 19 years. Since chemistry is often taken in the junior year of high school in America, the average student age of 17.3 at the time of this survey was expected. Most of the students plan to continue their education after high school either at a two-year college (9%) or a four-year university (77%). Surprisingly, none of the students indicated they planned to begin full time work or enlist in the military immediately after high school. Nevertheless, 52 students (almost 40%) reported that they worked outside of school hours. They worked an average of 17 hours each week, and one student claimed to work 40 hours each week! There was no correlation between the hours a student worked and the student's grade or ethnicity. Other students were also busy outside of classes. Almost 90% of the respondents indicate that beside jobs, they were also involved with extra-curricular activities such as drama, music, athletics or some other activity.

In school, the students generally do well. The average overall grade point average (GPA) of the students ranged from a low of 2.00 (C) to a high of 4.5 (A+) on a weighted 4.0 scale. The average GPA was 3.2 (B) and showed an approximate normal distribution. See Figure 3-4.
Figure 3-4. This figure shows the distribution of actual student GPA plotted against the student GPAs that would be expected if GPAs were normally distributed. As the actual student GPAs generally lie close to a line with slope of 1, one can consider GPA to be approximately normal in their distribution.

The students reported similar evaluations in the first semester of chemistry, although grades covered a slightly larger range. First semester chemistry grades ranged from a low of 1.00 (D) to a high of 4.50 (A^†). The average first semester grade was 3.4 (a high B) and, unlike GPA, showed a leftward skew. See Figure 3-5. Three students with very low first semester grades had dropped Ms. Lindsay's class before this study began. This could account for some of the skew in the grade distribution of the students in this study. While the effect is small, one could argue selection bias; however, as will be discussed, this narrowing of the population did not seem to effect the distribution of other "achievement" measures. The degree to which this student sample represents a standard American High School chemistry class will be addressed in the conclusion to this chapter.
**Figure 3-5.** This figure shows the distribution of actual student first semester chemistry grades plotted against the student first semester chemistry grades that would be expected if those grades were normally distributed. This figure indicates a slight left skew in first semester grades.

Almost all students reported they had completed two semesters of Algebra 1 (N=120) and the first semester of Geometry (N=121). One hundred fifteen of the students had completed both semesters of Biology. Although the average grades the students achieved in these classes is much lower than the group’s average chemistry grade, the inter-correlation between their grades in these classes and overall GPA are large and significant (see Figure 3-6). These correlations support Gardner’s contention that grades and GPA, in large part, measure the same, limited aspects of a student’s ability. (Gardner 1983) Furthermore, students who have excelled at school, as measured by the number of Advanced Placement classes taken, tend to have higher grades than those who have not taken any AP classes. The overwhelming majority of students in Ms. Lindsay’s class have taken no AP classes.
<table>
<thead>
<tr>
<th></th>
<th>GPA</th>
<th>Chem1</th>
<th>Bio1</th>
<th>Bio2</th>
<th>Algebra1</th>
<th>Algebra2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chem1</td>
<td>.673</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bio 1</td>
<td>.471</td>
<td>.398</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bio 2</td>
<td>.443</td>
<td>.426</td>
<td>.610</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algebra 1</td>
<td>.500</td>
<td>.502</td>
<td>.461</td>
<td>.357</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algebra 2</td>
<td>.564</td>
<td>.477</td>
<td>.334</td>
<td>.343</td>
<td>.696</td>
<td></td>
</tr>
<tr>
<td>Geom 1</td>
<td>.532</td>
<td>.505</td>
<td>.311</td>
<td>.379</td>
<td>.490</td>
<td>.443</td>
</tr>
</tbody>
</table>

**Figure 3-6.** This figure shows the bivariate inter-correlations between GPA and student reported grades for various high school math (first semester Algebra, second semester Algebra, and first semester Geometry), and science (first semester Biology, second semester Biology, and first semester chemistry) courses. Note that while the correlation between semester grade and GPA varies, the correlation between first and second semester grades in different subjects never falls below $r = .6$. This suggests that the manner by which an individual teacher determines grades and the way students perform is consistent from first to second semesters. All correlation coefficients are significant at $p < .001$.

The literature also suggests that one should be aware of a correlation between other variables that might suggest the presence of bias in assessments. While there is a slight, but significant correlation between ethnicity and overall GPA ($\chi^2= 10.04$, df = 3, $p = .018$), there is no significant correlation between ethnicity and first semester chemistry grade. Conversely, while there is no significant correlation between gender and overall GPA, there exists a significant correlation between gender and chemistry grade ($\chi^2= 9.6$, df = 3, $p = .02$) with female students achieving higher grades. Here again, this correlation is expected (see, for example, (AAUW 1992)). There is no significant correlation between SES and GPA or SES and first semester chemistry grade.

As discussed in Chapter 2, some suggest at least a marginal relationship between motivation, test scores and grades, therefore the student survey asked questions to explore this relationship. Overwhelmingly, the students liked both the class (84% said they like the class at least “some”) and their teacher (86% said they liked the teacher at least “quite a bit”). The vast majority of students (86%) also said that Ms. Lindsay’s evaluation of what they “know about chemistry is accurate.” Nevertheless, the students were almost equally divided when asked whether they found chemistry “frustrating.” More importantly, this variable was significantly correlated with GPA ($\chi^2= 15.5$, df = 3, $p = .001$), with first semester chemistry grade ($\chi^2= 21.9$, df = 3, $p < .001$), and with how well the student liked the class ($\chi^2= 23.1$, df = 3, $p < .001$).

The survey also investigated a student attitude closely related to student motivation, namely how much and what kind of effort the student felt was needed to successfully complete chemistry. While students report spending about 20 minutes each night, on average, “working on chemistry”, more than 64% agreed with the statement that they “don’t get a new concept…until I apply it in homework” and most (61%) said that they had to think about a concept for at least a day before understanding it. Surprisingly, only about half the students felt that laboratory work was important for understanding chemistry, and most students (60%) felt they could “memorize things for this class, take a test, and then forget about them” in order to receive a passing grade. None of these
variables had any significant correlation with first semester chemistry grade suggesting that these attitudes were held by both high achieving and lower achieving students alike.

There were no significant correlations between student attitudes about "problem solving" and any other measure.

3.2.1.2 Pre-treatment teacher survey and results
Prior to introducing IMMEXTM into the classroom, I also interviewed the students' high school chemistry teacher. This interview took place on February 10, 2000. The interview consisted of ten questions about each student in Ms. Lindsay's four chemistry classes, and the questions focused on teacher attitudinal variables that have been reputed to affect assessment validity. The actual questions asked during the interview are at Appendix 2.

Brewer and Kallick report research that suggests "significant groups of parents, particularly those of high school students and students who are not doing well in school, would like better information about how well their students are performing." They further suggest that parent involvement varies along demographic dimensions and the characteristics of the school their children attend. (Brewer and Kallick 1996) Given the relatively high SES of Lincoln High School, one might expect parent involvement to be high and that this involvement might have an effect on student assessment. Nevertheless, overall parent involvement with Ms. Lindsay was low. There were a total of 36 parental contacts during the first semester made by 23 parents. Most (14) of the parents made only one contact each, but one parent contacted Ms. Lindsay four times. The number of contacts correlated significantly with two other variables, the number of hours a student worked at their job each week (r=.323, p = .002), and first semester chemistry grade. This latter correlation remained significant even when the number of hours a student worked was controlled for. As might be expected, the correlation between parental contact and first semester chemistry grade was negative (r=-.219, p=.012).

More importantly, the relationship between teacher attitude about a student and that student's first semester grade was also explored. Ms. Lindsay indicated she enjoyed every student in her class, but that the last period of the day (5th period) was the most challenging. In fact, for whatever reason, the average grade for students in 5th period was lower than Ms. Lindsay's other periods. The differences, however, were not significant.

Overall, Ms. Lindsay felt that her female students "work harder" and are "more motivated." Nevertheless, while feeling male students were "more disruptive" she also felt that more disruptive students were also "more inquisitive". As might be expected, the students Ms. Lindsay classified as "more disruptive" and "more inquisitive" also spoke up more in class. None of these indicators, however, correlated with first semester chemistry grade.

As an indicative of the expectations she had formed for individual students, I asked Ms. Lindsay to rank each student within a class period by how well she expected them to do in the qualitative chemistry unit they would study later that semester. I gave no other specific indication of what she should use to rank the students. This was considered to be
an important pre-treatment predictor variable. Ms. Lindsay’s ranking is significantly correlated with GPA (r = -.577, p < .001) and has an even larger correlation with first semester chemistry grade (r = -.729, p < .001). Again, given the large significant correlations between first and second semester grades suggested in Figure 3-6 above and other literature, these findings were not surprising. In fact, Ms. Lindsay later confided that she had based her rankings largely on the student’s first semester chemistry grade. Nevertheless, grades are not the only measures that correlate with her ranking. The student’s rank is also very slightly correlated with Ms. Lindsay’s perception of each student’s motivation ($\chi^2=12$, df = 6, p < .06), but this correlation does not meet the threshold of significance required in this study (p < .05). However, when trying to develop a model to predict her ranking, adding GPA or motivation to first semester grades did not significantly improve the model.

Since many of the conclusions in this study are based on student responses, I also asked Ms. Lindsay for the students’ first semester chemistry grades. I then correlated her response with the response given by each student. Student and teacher responses were overwhelming identical (r = .955, p < .001), and differences were confined to “rounding” (e.g. a student would report a first semester grade of B and the teacher would indicate B').

3.2.1.3 Pre-treatment IMMEX™ problem
As a final activity in the pre-treatment phase, the students were assigned the task of completing an IMMEX™ problem unrelated to the IMMEX™ problems they would work with during the treatment phase of the research. This IMMEX™ problem, How dense are you?, presents students the problem of determining the volume of an unknown given its identity and mass. While straightforward and simple to solve in that all but one of the 134 students eventually solved one of the clones, it exposed the students to the IMMEX™ software and the computer technology that most (95%) had not experienced before this study. As such, we attempted to remove any extraneous IMMEX™ or technology related variables such as unfamiliarity with what an IMMEX™ problem was asking or how to navigate through the problem space from the results in the treatment phase of the study. This pre-treatment also allowed me to ensure that the technology was working correctly.

3.2.2 Treatment
The students began using the IMMEX™ problem Hazmat Holiday Special (Hazmat) for learning and practice on May 1, 2000. The problem was available to them over the World Wide Web 24 hours a day until May 9th. This corresponded with the middle of their qualitative chemistry unit (Unit 12).

---

4 The significance level of .05 was chosen largely due to convention; however, I also felt that with the relatively small sample size of this study (N=134), that a more stringent threshold might be difficult to achieve and might disqualify many otherwise important relationships between variables.
The *Hazmat* problem was developed by Ms. Lindsay and four other teachers. In this problem the student is told there has been an earthquake which has caused a number of chemicals, some of which may be hazardous, to fall off the stockroom shelf. As the labels are not with the spilled chemicals and time is of the essence, the school has hired some of its chemistry students to identify the spilled chemicals. There are 12 physical or chemical tests the students can conduct on the unknown substance and 8 general reference items students may review in their attempt to identify the unknown substance. The students were given class time to complete at least two clones of this problem. In addition, they were provided "worksheets" developed by Ms. Lindsay. A copy of this worksheet is located at Appendix 3. The students were told that these worksheets would be graded along with their performances as recorded by the computer. In addition, Ms. Lindsay indicated to her classes that the students would have one of these problems on their end of unit test.

The students each completed almost 5 clones of *Hazmat*, on average, during this practice and learning period. Generally, each student was successful on about half (2.3) of the clones attempted. The success rates on these problems were not significantly different between class periods as shown in Figure 3-7. As Ms. Lindsay expected, the mean performance for Period 5 was the lowest of all the means and it had the lowest second quartile; however, its upper third and fourth quartiles ranked with the highest periods.

![Box and whisker plots for Ms. Lindsay's four periods of chemistry (Periods 2, 3, 4 and 5). The means are not significantly different, although the variances between periods are quite different.](image-url)

**Figure 3-7.**
The students sat for a pencil and paper unit test in the middle of May. A copy of this test is at Appendix 4. Besides testing ionic and covalent bonding, chemical equations and molecular structures, the test was designed to evaluate each student's understanding of concepts previously discussed during the year. The pencil and paper test consisted of computational, fill-in the blank, matching and multiple choice questions to assess student learning. The distribution of grades on the Chapter 12 test is shown in Figure 3-8.

![Normal P-P Plot of Unit 12 Test](image)

**Figure 3-8.** This figure shows the distribution of actual student grades on their Chapter 12 (qualitative chemistry) test plotted against the student test scores that would be expected if those scores were normally distributed. As the actual student scores generally lie close to a line with slope of 1, one can consider these test scores to be approximately normal in their distribution. The variance in overall scores is, however, smaller than normal.

In addition, the students ran at lest one *Hazmat* problem on May 16, 2000 as part of their Unit 12 assessment. Because the students ran so few problems we can conclude little from this data in isolation, other than noticing what appears to be a large difference in period three’s performance. The 31 students in period 3 solved about 45% of the problems they, as a group, attempted. This was significantly different from period 4 where students averaged a 70% solve rate and period 2 where the students averaged a 77% solve rate. The difference between period 3 and period 5 (where the students averaged a 61% solve rate) was not large enough to be statistically different. Nevertheless, this outcome was quite unexpected by Ms. Lindsay.
3.2.3 Post-treatment

Shortly after the unit test, Ms. Lindsay’s four classes took the State Test of Academic Readiness (STAR). This test consists primarily of the Stanford Achievement Tests – version 9 (SAT-9). The SAT-9 reports three scores that we discuss here: a math score, a science score, and a reading score.

Although the SAT-9 was not designed to reflect the current state standards or frameworks, it was expected that the SAT-9 would correlate with academic grades for many of the reasons cited earlier. In particular, since these tests seem to be an important barometer for the schools and improving these scores appears to be an important goal, it was expected more classroom time would be devoted to teaching to this test. I also expected a direct relationship between gender and both SAT-9 math and science scores. Nevertheless, the correlation between grade measures and the three sections of the SAT-9 yielded some surprises.

The inter-correlation between the three sections of the SAT-9, for Ms. Lindsay’s students, is strong and significant. This can be seen in Figure 3-9.

<table>
<thead>
<tr>
<th></th>
<th>SAT-9 reading</th>
<th>SAT-9 math</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT-9 math</td>
<td>.336</td>
<td></td>
</tr>
<tr>
<td>SAT-9 science</td>
<td>.642</td>
<td>.540</td>
</tr>
</tbody>
</table>

**Figure 3-9.** This figure shows the correlation coefficients between SAT-9 math, science and reading scores for Ms. Lindsay’s students. All coefficients are significant at p < .001.

On average, the students in Ms. Lindsay’s classes scored at the 60th percentile on SAT-9 science, at the 67th percentile on SAT-9 math, and near the 56th percentile on SAT-9 reading. The tests indicate a significant correlation with student GPA. While SAT-9 reading shows the strongest correlation (r = .465, p < .001), SAT-9 math and science also show significant correlations with GPA (r = .288, p = .001) and (r = .291, p = .001), respectively.

All three SAT-9 tests also show significant correlations with first semester chemistry grades (r = .363 for science, r = .356 for math, and r = .414 for reading, all at p < .001). While correlations between grades and each of these tests was expected, that the largest correlation would be between grades and reading was unexpected. In fact, the student’s SAT-9 reading scores correlate more strongly than either of the other SAT-9 tests with almost every math and science grade the students have received. In addition, all three SAT-9 test scores are significantly correlated with Ms. Lindsay’s predicted ranking of how her students would do in the qualitative chemistry unit made in the pre-treatment phase of this research.

SAT-9 math shows a significant correlation with gender, as expected (r = .252, p = .004; \( \chi^2 = 8.947, p = .03 \)), but, surprisingly, the science test does not show a significant correlation. SAT-9 reading shows a strong significant correlation with Ethnicity...
\( \chi^2 = 20.83, \text{df} = 3, p < .001 \). White students overwhelmingly score above the mean on this test, while non-white students overwhelmingly score below the mean.

Of the SAT-9's three tests, it seems odd that a student's SAT-9 science has the lowest correlation with first semester chemistry grade and that the SAT-9 reading score has the best correlation with this grade. One could argue that the chemistry grade might be measuring knowledge outside the domain loosely referred to as science. Nevertheless, given that SAT-9 reading (not science) scores also correlate best with other science grades, and that the same observation can be made in the mathematics domain, such an argument seems implausible.

After their unit on qualitative chemistry and taking state tests, the students proceeded onto another topic in the course of instruction. During this unit, they continued to practice their qualitative analysis skills using an IMMEX\textsuperscript{TM} problem similar to Hazmat. The problem, called Desperately Seeking Solution (DSS), nearly quadruples the number of chemical tests a student can run and slightly reduces the amount of reference information available to the students. As before, Ms. Lindsay provided her students worksheets for this problem. This worksheet is at Appendix 5. It should be noted that DSS was originally designed by Ms. Lindsay and another teacher to test more advanced chemistry knowledge than taught to college preparatory chemistry students. A description of how this problem was adapted for Ms. Lindsay's class will be described in the next chapter.

Ms. Lindsay's students practiced DSS from June 1 - 8, 2000. As with Hazmat, they were given class time to complete at least two problems. Here again, they were told to complete the provided worksheets each time they attempted a problem. On average, each student completed almost three (2.6) clones of DSS during the week it was available to them for practice. This number represents just half the average number of Hazmat clones these students attempted. As was previously the case, no students completed any clones outside of class time even though they were available to them. On average, the students successfully solved slightly more than one DSS clone each. The mean solve rates between periods were not significantly different, as seen in Figure 3-10.
**Figure 3-10.** This figure shows box and whisker plots for Ms. Lindsay's four periods of chemistry (Periods 2, 3, 4 and 5). The means are not significantly different, although the variances between periods are quite different. As was the case in Hazmat, period 4 again had the largest variance and period 3 had the lowest maximum solve rate as well as the lowest median solve rate. There are two outliers. One, in period 2, answered no DSS clones correctly in practice; the other, in period 3, correctly answered all the DSS clones attempted.

Ms. Lindsay's students also used this IMMEX™ problem for assessment purposes. Because each student ran, on average, only one problem, we can conclude little from the data by itself except to note that there were not significant differences between the median solve rates of the four periods.

The students sat for their final written chemistry assessment from June 13 - 15, 2000. The final included many items similar to those the students had seen on the Unit 12 test in May. The final also included a computation of pH, the structural properties of molecules, and concepts involving equilibrium. As before, the pencil and paper section of the test used computation, fill-in the blank, matching and multiple choice to assess student learning. A copy of the final test is at Appendix 6. In addition, each student was asked to complete at least one clone of DSS as part of the assessment. Ms. Lindsay included these performances in her final grade for each student.

I administered a follow-up student survey to each student before they took the final exam. Seventy-seven percent of the students responding (N = 128) indicated that they had used at least "some" knowledge learned during the school year to solve IMMEX™ problems, and there was a strong significant correlation between how much a student reported using learned knowledge and how much they liked using IMMEX™ ($\chi^2 = 21.5$, df = 4, p < .001).
On the other hand, seventy-six percent of the students responding reported they only guess at a solution “some” (34%), “very little” (27%) or “not at all” (15%). As was expected, students who report guessing more often at an IMMEX™ solution also reported that they were less likely to connect the knowledge learned in this class with IMMEX™ problem solving ($\chi^2 = 18.94$, df = 6, p = .004) and were less likely to feel IMMEX™ improved their problem solving skills ($\chi^2 = 15.26$, df = 4, p = .004). There exists a strong and significant relationship between the frequency students report guessing at IMMEX™ problems and their GPA ($\chi^2 = 17.3$, df = 6, p = .008). In this case the more a student reports guessing, the more likely the student earned a lower GPA.

Similarly, 29% of the student respondents claimed they were able to guess correctly “quite a bit” (19%) or “a lot” (10%) when solving IMMEX™ problems. Here again, there is a significant, negative correlation between this variable and how much a student reports liking IMMEX™; however, there is no significant correlation between this measure and any of the grade measures. This was somewhat unexpected, especially since the correlation between those students who report guessing and those reporting guessing correctly is significant ($\chi^2 = 30.22$, df = 4, p < .001).

A full seventy-six percent of the students report that they believe IMMEX™ improves their problem solving skills and, as implied above, these students were much more likely to like using IMMEX™ and less likely to report guessing at IMMEX™ problem solutions.

As part of the final survey, the students were also asked to grade themselves on “how well you know qualitative chemistry.” This self-evaluation showed a significant correlation with GPA ($r = .248$, p < .001), first semester chemistry grade ($r = .281$, p < .001), and Unit 12 test grade ($r = .273$, p < .001).

Neither the math nor verbal sections of a student’s most recent SAT score was significantly correlated with any student grade measure. It is, however, interesting to note that both math and verbal scores were correlated with SAT-9 reading. On the other hand, only SAT math score was significantly correlated with both SAT-9 math, and SAT-9 science.

I held a final meeting with Ms. Lindsay at her school on Friday, July 7, 2000. The two important observations she shared at that meeting will be discussed in subsequent chapters.

3.3 Conclusions
The IMMEX™ software outlined above has many of the characteristics suggested by the literature as necessary to truly assess student learning and understanding. Based on these characteristics and the way it was employed in this study we should expect that the inferences made from the data generated by IMMEX™ will have greater validity than the inferences we make about student learning and understanding using the data generated by current pencil and paper assessments. As an authoring platform, IMMEX™ seems to
allow teachers to present problems to their students that fit the content of their instructional curriculum and the context of the instructional setting. In addition, it permits students to work on tasks that their teacher feels are authentic. Messick argues that the nature of the construct (understanding in this case) must guide the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. In the interest of reality testing and generalizability he also argues it would be desirable if the test were related to real-world behavior. (Mislevy, Steinberg et al. 1999) As we have seen, while pencil and paper tests have their useful applications, these tools are often used in ways their designers never envisioned or actually caution against. Because such tests are not “adaptable,” they must be applied in a situation as written. The authoring module of IMMEX™ overcomes this limitation. Teacher authors may adapt or create a problem based on the construct or criterion they are interested in measuring and the content in the instructional curriculum. Although this will be addressed more fully in the next chapter, it seems the student’s teacher is much more likely to know the students and their curricular requirement than is a test maker or software designer many miles removed.

That IMMEX™-like problems allow for the application of knowledge and so measure a student’s ability to perform better than similar pencil and paper tasks is argued for by Webb, et al. She and her colleagues suggest that pencil and paper tasks may only be considered interchangeable with “hands on tasks” if sufficiently large numbers of pencil and paper tasks, raters and occasions are used to reduce measurement error to near zero. (Webb, Schlackman et al. 2000) As we have seen, limited class time and the increasing constraints on that time, including using instructional time to teach to new tests, make more time for new pencil and paper tasks unlikely. Furthermore, if the pencil and paper tasks bear little resemblance to what is expected in the world beyond the walls of the classroom, they may not be preparing the student to perform well in the future. In this case, “teaching to the test” becomes even more problematic. Authentic problems that can be delivered anytime via the World Wide Web, and therefore can be practiced outside of limited class time, can skirt this limitation. Conceivably, a student could prepare for such assessments as a lawyer prepares for trial arguments, or other professions make ready to practice their craft. This form of “teaching to the test” is learning and anticipating the unexpected, not memorizing the expected.

While one might argue that pencil and paper tasks can also be accomplished outside the narrow limits of class time, unlike IMMEX™, tasks accomplished in this way are not subject to the type of inspection afforded by IMMEX™ technology. This software not only allows educators to see what a student answered, but allows all stakeholders to actually see how the student went about arriving at that answer and what information they used to arrive at answer. If Brady’s concept of educating for understanding is to be believed, an ability to see these types of student actions is critical to an educator’s work. Brady argues that education for understanding is primarily concerned with the construction, organization and elaboration of conceptual frameworks composed of related ideas. Furthermore, she believes that instruction not aimed at creating or clarifying a concept’s relationship within such frameworks ignores the mind’s usual approach to processing information. (Brady 1989) Numerous other brain and cognitive scientists
share Brady's conceptualization of how the mind processes information (see, for example, (Michalski and Stepp 1983; Mann and Jepson 1993; Richards, Jepson et al. 1996)).

While the technology component used in this study must afford educators certain assessment capabilities they currently lack in order to improve the accuracy of the inferences they make from assessment data, technology alone is insufficient to demonstrate that the resulting inferences are any more accurate. A suitable methodology must demonstrate that resulting inferences have, in fact, improved.

This chapter described the quasi-experimental, interrupted time-series approach used in the present research. This methodology is particularly well suited to educational studies in that it allows research to take place in the natural context of the classroom. Nevertheless, such a methodology requires that a sound baseline be established so that post-treatment comparisons can be made from which to infer the effects of treatment. Furthermore, a detailed understanding of the population involved in the study suggests the degree to which findings are applicable to a wider population.

While the students described in this chapter are overwhelmingly white and belong to families of upper-middle class socioeconomic status that is not atypical of the U.S. suburbs, and Ms. Lindsay's students seem, by most measures, to be representative of American suburban, high school students. This is especially noticeable in the baseline academic performance measures described. The student's overall weighted GPAs are roughly normally distributed and their first semester chemistry grades, although skewed left, correlate with other math and science grades in the same way these grades correlate with one another. Therefore, while chemistry grades may be higher, they seem proportionally higher for all students in the class. As the literature and these findings suggest, grading is largely a personalized activity of individual teachers. Nevertheless, it seems Ms. Lindsay does consider prior grade an important predictor of future student performance. Significantly however, the students felt that by memorizing concepts, they would earn "acceptable" grades. On the other hand, students who resorted to guessing in order to complete the class generally had lower overall GPAs but not lower first semester chemistry grades. The literature suggests that female students typically earn higher grades than their male classmates and that trend is present in this population. As is characteristic of most schools, students in this cohort who are less "frustrated" with the subject, also seem to do better in this class.

As was expected, these students tend to exhibit typical "performances" on standardized tests, although the gender bias against females reported in the literature is evidently not present for this population. Nevertheless, non-white students appear to achieve significantly lower scores on the reading portion of the state test of academic readiness than their white counterparts. And, surprisingly, it is this test, not the math or science portions, that correlates most strongly with math and science grades. Finally, as reported in the literature, there exists a strong correlation between all grade measures and standardized tests of "achievement," including the student's own evaluation of how well they know the material presented in the unit on qualitative chemistry.
Consequently, by most measures, this student population appears representative of the larger population of suburban high school students taking chemistry. As such, I suspect the findings presented here are generalizable to the wider population of students taking chemistry at suburban U.S. high schools. While I believe that future research will demonstrate these findings apply in urban high school settings and to other disciplines as well, the ethnic and SES characteristics of this student cohort makes extrapolating these results to urban schools tenuous. Furthermore, since the research suggests we were able to uncover general problem solving strategies, I believe these results will also generalize to other disciplines of knowledge.

Given this baseline, the next three chapters detail how our present inferences of student learning and understanding can be improved. Since content dissimilarities seemingly account for many of the invalid inferences previously discussed, we now address how this technology can improve this aspect of assessment accuracy.
Chapter 4. Improving inferences of understanding by testing what was taught

A point made repeatedly in the previous chapters is that one of the biggest threats to accurate assessments in the United States educational system today involves the mismatch between the content of an assessment and the content of the curriculum that those being assessed experienced. This problem exists whether the test is delivered using pencil and paper or with modern computer technology. The content tested by most, if not all, present day large scale assessments is not required to, and so probably does not, match the content of the curriculum most students experience. There is no requirement for tests to match curricular content both because the associations that promulgate assessment standards like the American Psychological Association have no enforcement authority and because alignment is not necessary in all situations.

In actuality it is not the misalignment between test content and the content of a specific instructional curriculum that yields inferences of questionable accuracy. In fact, depending on the inferences to be drawn, a test need not, in every case, match the content of a particular curriculum. As previously mentioned, a test designed to sort or rank order students can be, in fact often is, designed around the questions most likely to discriminate students regardless of formal education. Question items in these tests are chosen not because they match a curricular content, but because they discriminate students well. Test makers often argue that a question’s ability to discriminate students is based on some latent ability, life long experience, or another similar construct. Such a test, however, is unlikely to measure specific content knowledge in a way necessary to be useful as a learning achievement test. Consequently, inferences made about student achievement from data generated by tests not designed for such a use are probably inaccurate. Nevertheless, some suggest even these non-achievement tests must demonstrate alignment with an identifiable content area in order to produce accurate results. As Bloom observes, “content validity and not item statistics should be the prime consideration in selecting items, regardless of whether one wishes to make normalized or criterion referenced inferences about performance.” (Bloom, Madaus et al. 1981) So while different tests might sample different content areas, they must still explicate what content area they sample from and, therefore, over what content area inferences based on the resulting data are valid. Consequently, it is not the test that is valid or invalid, rather it is the inferences made from the test data that should be so classified. This is true of “Intelligence” tests or other so called “aptitude” tests for which test makers claim there is no specific “curriculum.” It is also true for “achievement” tests.

As previously described, even though Binet specifically excluded the taught curriculum from his test of student ability in order to avoid merely testing the differential education children had been exposed to, he specifically included knowledge that test-takers should have gained by life experience. However, even Binet was loath to say that this measured native intelligence or, as we have seen, a specific amount of demonstrated intelligence. He was also very clear that his test did not measure student achievement, nor did he infer that students or their schools were “under achievers” when students performed poorly on his test. Rather, Binet was attempting to identify students who, because they lacked the ability to learn from “normal” experience in late 19th century France, might benefit from
other forms of instruction. Binet's test sampled a very specific cultural "curriculum." Unfortunately the results of misapplying Binet's methods in the United States, are well documented. (Gould 1996) More lamentable is that this misapplication continues, especially in American education, to this day. Tests of life experience might allow accurate inferences about whether a student has the background necessary to make sense of similar experiences, but they specifically cannot generate the data necessary to make inferences about a student's learning and understanding in a specific content area. Unfortunately, as the data generated by a single test is increasingly used as a basis for a growing number of different inferences, the threat of content invalidity grows.

In order to make accurate inferences about student achievement, one must minimize the problem of content invalidity. Accordingly, a test or assessment must represent the objectives or skills about which one wishes to make inferences. Previously, I have discussed international and national tests that purport to measure student achievement or ability in math and science. I have reviewed the claims some have used the results of these tests to justify. I have also demonstrated that, given the content of the test and how that content was defined, conclusions about how well American schools perform and how well students have achieved the standards set for them cannot be justified using these tests. This is primarily because these tests were not designed in light of the standards schools taught to. Surprisingly, even some state tests, like the SAT-9 in California, were not constructed against a state standard or curriculum. It should not be surprising then that when students are taught one curriculum and a state, national, or international test measures the same students against some other standard, different data results.

Merely changing the content of the curriculum to match the content of an assessment does not seem to resolve the disparity between curricular and test content in an acceptable manner. The literature suggests that when classroom teachers have no control over the content measured by a high-stakes test, the curriculum will inevitably change to cover that content. However, it also suggests that the resulting content will not be taught for student learning and understanding. Rather, when the curriculum changes to achieve high scores on high stakes, standardized tests, inevitably students are instructed how to do well on the test, not to understand the content. (Resnick and Resnick 1990)

Besides the curricular content, the 1985 Standards further suggest that test formats should also match the manner in which students were instructed (APA 1985), and Messick argues that "the nature of the curriculum or the actual classroom experiences...serve to delimit the domain." (Messick 1989) For example, according to these guidelines, it would be inappropriate to assess students who had studied a topic using hands-on activities, with a multiple choice test. In fact, Mosier suggests that, even if they seem to get at the same domain, two such test methods can yield results that are totally uncorrelated. (Messick 1989) Similarly, Webb et al. argues that pencil and paper tasks are not always interchangeable with hands-on tasks. (Webb, Schlackman et al. 2000) This research would suggest, as does the research of others previously cited, that multiple forms of learning and assessment might be necessary to instruct and test student learning and understanding. At present this seems only a fanciful dream given our over
reliance on pencil and paper, multiple choice tests in America, and our difficulty measuring the congruity between instruction and assessment.

Because most assessments must cover broad curricular domains in limited time, they typically can only sample a small percentage of the entire universe of possible questions or tasks in that domain. Therefore, to achieve the necessary content validity an assessment must demonstrate it only tests within specified domain boundaries (i.e. it is relevant) and that it samples across the entire domain space about which one desires to make inferences (i.e. it is representative). In addition, as suggested above, it should do this using multiple formats.

Against this background, Messick and others suggest that "subject matter experts" or teachers should decide if a test appropriately samples a given content domain. (Cronbach 1971; Messick 1989) However, given the great diversity of content taught in schools, it seems untenable that such an evaluation of a single test would yield consistent results. It is also unrealistic to expect that, given current standards, all "experts" would agree on how various content should be weighted. Burns, in fact, addresses this very point when arguing that, "a way must be found to move the process from the quantitative-face-validity...to a methodology that could correct for overly enthusiastic professionals and / or the apparent positive bias of "subject matter experts"." (Burns 1996) Furthermore, while "subject matter experts" could conceivably evaluate the match between published standards and a large scale test at the state level, they could not evaluate the actual experiences of students within all of the state's classrooms. Realistically then, much of the responsibility for evaluating the typical tests encountered by students seems to vest at a more local level.

More than anyone else, the teacher knows the content she/he taught to meet a specific standard, the weight that content was given in the curriculum, and the context in which it was taught. To a nearly identical extent, so do the teacher’s students. The teacher and students also are more aware of the "prior" experience students bring to the classroom than are distant test-makers and "subject matter" experts. Effective teachers and tutors use these "cues" to build effective curriculum and help the student understand concepts within a particular domain. (Lesh and Kelly 1996) In fact, even brain and cognitive science research now suggests that prior experience is critical to "making sense" of the world. (Richards, Jepson et al. 1996) Given such a contribution, it seems logical to consider prior experience when evaluating the match between both instructional and curricular content and the content on an assessment. In other words, depending on the previous experiences of the students, presenting an instructional video to two classrooms and then testing them about the content of that video, for example, represents "equal", but probably not equitable instruction and assessment. Cultural considerations in international tests are a special case of this consideration and were discussed in more detail in Chapter 2. For these reasons, among others, Wiggins concludes that the assessment of student achievement must be made at the local level. In addition, he also points out that the very root of the word "assess" means to "sit with" or sit beside. (Wiggins 1993) Likewise, the students themselves should not be overlooked as "experts"
on what a curriculum included and, more importantly, what and how they were required to learn that curriculum.

I propose that IMMEX™ technology offers American education the opportunity to improve the alignment between curriculum and assessment content over current, widespread testing practices. As will be discussed, this technology allows “experts” at all levels to judge the match between instructional and assessment content, and allows teachers to adapt the content of both practice and assessment to the context she requires. Furthermore, the flexibility of IMMEX™ allows practice forms and assessment forms to be similar or identical, and permits tasks to be presented in a way that minimizes the effect of English literacy as a confounding variable in learning and assessment.

4.1 Allowing subject matter experts to evaluate content validity

IMMEX™ problems are usually designed by actual teachers for use in an on-going curriculum. Some problems have also been designed by “subject matter experts” such as medical doctors for patient education or researchers as part of a study. As a standard part of this design process, new problems are shown to other teachers or experts. Furthermore, the problems are generally run by one or more groups of students before they become widely distributed.

When problem testing is complete, the problem is posted on the IMMEX™ web site for public use and review. The results from actual student trials are also studied by teachers and researchers. Perhaps the most intensive subject matter review of particular IMMEX™ problems comes when they are presented as part of research in papers or at conferences. For example, in addition to being used by 9 teachers and 2000 students, Hazmat has been presented at numerous conferences and described in numerous peer reviewed papers over the last decade. Additionally, since it is available on the World Wide Web, the Hazmat problem is accessible to both subject matter experts, and to the wider public as well. Similarly, Desperately Seeking Solution has been used by 14 teachers and by almost 3000 students, has been studied extensively by researchers, and has been the subject of numerous refereed articles.

While the subject matter of the Hazmat problem has remained largely unchanged since its inception, outside experts have recommended some changes. Since the problems rely greatly on the interpretation of visual clues, some have suggested that the problems would be inappropriate for visually impaired students. The IMMEX™ staff is studying how to accommodate these changes. Another teacher has suggested the addition of “smell” as a possible menu item. While delivering an actual odor is not technologically possible at present, teachers may add these and other written descriptions as new menu items whenever they see fit.

4.2 Allowing the teacher to function as content expert

The ability of teachers to modify IMMEX™ problems to fit the context and instructional curriculum of a particular group of students seem to contribute greatly to improving the content validity of assessment in U.S. classrooms using this technology.
As one of the problem authors, Ms. Lindsay developed *Hazmat* and *Desperately Seeking Solution* to integrate with the curriculum she teaches. Arguably, however, the abilities of students and the curriculum taught vary from year to year. In fact, during the period of this study Ms. Lindsay had not yet covered the material necessary to perform all the clones in either of the two problems.

In both *Hazmat* and *Desperately Seeking Solution*, students are asked to identify compounds consisting of two parts – a positive ion (cation) and a negative ion (anion). The cations fall into two broad categories. The easiest cations to identify are those that color a flame they are placed into. The more difficult cations don’t change a flame’s color and, therefore, require more complex tests to identify. Ms. Lindsay had not instructed the students in this second group of tests, consequently it would have been inappropriate to assess student ability to use such tests. In its standard format, the content of the IMMEX™ problem would not have matched the curricular content of Ms. Lindsay’s class. With prepackaged software or pre-made pencil and paper tests, adapting the assessment to what was actually implemented in the classroom would have been difficult, if not impossible. Because IMMEX™ allows teachers to modify the problem to suit her needs, Ms. Lindsay was able to easily modify the problem to align with her curricular content.

Ms. Lindsay removed all the clones of *Hazmat* that produced uninformative flame tests. Consequently the students were only solving for unknowns that involved six cations (Sodium, Calcium, Barium, Copper, Potassium, and Strontium). There are five possible anions (Chloride, Sulfate, Hydroxide, Nitrate, and Carbonate) in the *Hazmat* problem. The various combinations of cations and anions produced a possible thirty unknowns of which twenty three clones are actually implemented. This method of generating “item forms” is similar to the methodology suggested by Osburn, Hively, Patterson, and Page. (Messick 1989)

As described in Chapter 3, Ms. Lindsay also provided the students worksheets (see Appendix 3) in order to further structure their thinking and problem solving while using IMMEX™. She used these notes to assess student performance and understanding in both practice and assessment.

Ms. Lindsay confided in our final interview that she believed IMMEX™ allowed her to measure student achievement in a way standard pencil and paper tests, even those tests she created, did not allow. She used the following example. On pencil and paper tests, a question might require students to predict the products and results of the following chemical reaction:

\[
\text{AgNO}_3(\text{aq}) + \text{NaCl}(\text{aq}) \rightarrow ?
\]

Stated in this manner, solving the problem requires the student demonstrate basic algorithmic or memorized knowledge. In this case the student would remember to rearrange the positive and negative ions, then check the solubility table. The teacher would then give the student credit for showing Ag combines with Cl and Na with NO₃. In
addition, the student would receive credit for determining that AgCl was not soluble in water and would precipitate, while NaNO₃ would remain dissolved. As Ms. Lindsay pointed out, she could not determine if the students understood this, or merely followed the algorithm when faced with such problems on her standard tests (see Appendices 4 and 6). With IMMEX™ however, the student must not only hypothesize which combinations of compounds will produce which results, but s)he must also effectively reason from the right side of the above equation back to the left. In addition, the student must include the possible combinations that could yield an observed result and eliminate those combinations that would not. It is this synthesis of knowledge that Ms. Lindsay finds difficult to assess using pencil and paper. But it is also just this type of learning Bloom felt it so critical that we assess in education. (Bloom 1956) In fact, evidence suggests this type of knowledge is even difficult to measure when students are required to perform and write up “typical” lab experiments. Not only do written labs often require students follow a recipe, but students often filter written reports to include only what they think the teacher desires or what is “correct.” (Raizen 1990) Furthermore, in such reports poor written communication skills can confound an accurate assessment of a student’s ability in qualitative chemistry. With IMMEX™, all aspects of a performance can actually be “seen” by the teacher.

The importance of this kind of evaluation was made even more poignantly by several students during this study. The students generally had particular difficulty solving Hazmat problems. Nevertheless, all the students had correctly solved at least one clone and had written down all the test results in their notes. When they sat for their assessment, they merely duplicated their practice note sheet as an assessment result. In all cases, computer records indicated they had not performed an assessment problem. Furthermore, the identity of the unknown they “solved for” was not a possible unknown during the assessment period. The students probably did not understand the desired content, but they wanted their teacher to believe they did.

IMMEX™ technology also allowed Ms. Lindsay to ensure that assessment matched exactly what the students practiced. During practice, the students performed tasks that involved all the possible cations and all possible anions, but not all possible combinations of the two. So while the students may have practiced with the unknowns Sodium Chloride and Potassium Carbonate, they would only have seen Potassium Chloride on the assessment. Because one can use the same procedure to identify the Chloride ion regardless of which cation it is presented with, the assessment and practice problems are, for all practical purposes, identical.

Ms. Lindsay covered the concepts in Figure 4-1 during the unit on qualitative chemistry. The students practiced and were assessed on each concept as indicated in the figure. In each case, the method of assessment was identical to the method students used to learn and practice that concept in the unit.
<table>
<thead>
<tr>
<th>Topic covered or reviewed in Chapter 12</th>
<th>Assessed using Pencil and Paper</th>
<th>Assessed using IMMEX™ technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific Notation</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Significant Figures</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Dimensional Analysis</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Atomic Structure</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Molecular Structure</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Molecular Formulas</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Chemical Bonding</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Chemical Properties of substances</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Balancing Equations</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Ionic Equations</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Types of reactions</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Prediction of Products</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Application of Chemical concepts</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Physical Properties of substances</td>
<td>✅</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4-1.** This figure shows the concepts the students in Ms. Lindsay’s class learned or reviewed during the qualitative chemistry unit. The students used the method indicated to practice a specific concept and were assessed on that concept using the same method.

**4.3 The student as expert**

Perhaps those with the greatest stakes in the educational process and those most intimately involved in the process are the students who sit in our classrooms each day. As such, a good indicator that tests actually cover only the specific curriculum that the students were exposed to and that assessments were appropriately weighted across the curriculum would be student attitude about how they are assessed. Surprisingly, student opinions about the alignment between instruction and assessment are rarely sought.

As discussed in Chapter 3, during the pre-treatment phase of this research, students were overwhelming positive that their teacher accurately evaluated their knowledge of chemistry. After using IMMEX™ software, the students were also asked how often they used knowledge learned during the school year to solve the problems this software presented to them. More than 75% of the students indicated they used at least some knowledge learned in this class to solve IMMEX™ problems. This suggests a good alignment not only with the material covered in the qualitative chemistry unit, but with Ms. Lindsay’s chemistry course overall.

In addition to overt student survey responses, the tools of modern test theory offer the ability to compare student performance on individual *Hazmat* clones both to the student’s performance on *Hazmat* overall and to compare the student’s performance between
Hazmat clones involving different anions. One would logically expect that without exposure to a certain topic, students of all ability would do poorly on one or more problems of a given type. Therefore, if a problem from outside the instructional area was introduced to students on the assessment, student performance on this problem type would be uncorrelated with overall student ability. Similarly, if student ability on one type of problem correlated highly with ability on another type of problem, the two problems might be measuring the same aspect of the curriculum. Before discussing the actual results of this analysis, the analysis tool itself (Item Response Theory) is reviewed.

4.3.1 Item response theory

In classical test theory, the difficulty of a test or a single test item is determined by the ability of the particular students taking the test. Therefore, if a group of students gets a particular test question correct only 25% of the time, the question might be considered a “difficult” question. Ultimately, however, how students perform on the test question or test as a whole depends on what group of students one chooses to test. Consequently, great efforts are made, with varying degrees of success, to find a representative, “normal” sample against which to “calibrate” the test. Modern test theory, which was developed and refined by Rasch and others since the 1940’s, attempts to focus not on the test, but on the individual items comprising the test. In addition, it attempts to measure some sort of latent ability in the student that is related to her or his ability to answer the posed question. Furthermore, this Item Response Theory (IRT) assumes that students with more of this latent ability have a greater likelihood of answering an item correctly than students with less ability. In the late 1950’s researchers suggested using a logistic regression curve to model the probability a student would correctly answer an item, given that student’s ability. (Baker 1985) In its most elaborate form, the probability a student solves the posed problem correctly is:

\[
Pr(\text{correct}) = \frac{1 + ce^{-a(\theta-b)}}{1 + e^{-a(\theta-b)}}
\]

where

- \( e \) is the natural number,
- \( a \) is a measure of the item’s ability to discriminate,
- \( b \) is the difficulty parameter of the item,
- \( c \) is the probability a student can guess correctly, and
- \( \theta \) is the ability level of the student.

Rasch suggested the so called “single parameter” model in the 1960’s as an acceptable alternative to this more complex model. (Baker 1985) Essentially, the Rasch model does not penalize students for guessing and assumes that the ability of a problem to discriminate students likely to answer a question correctly or incorrectly is fixed. As such \( c = 0 \) and \( a = 1 \). These assumptions, however, leave the essential part of the theory in tact in that it allows the difficulty of individual items to vary. The “Rasch” model then becomes:
Pr(correct) = \frac{1}{1 + e^{-(\theta-b)}}

The curve this function describes is shown in Figure 4-2.

**Figure 4-2.** This is the logistics curve proposed by Rasch as a model for the probability a student of ability \( \theta \) will correctly answer a specific test item. When the student's ability equals the difficulty parameter of the problem (i.e. the "deviation" between student ability and problem difficulty is zero), the likelihood the student answers correctly is shown by this curve to be 50%.

In the Rasch model, the term \((\theta - b)\) is termed the logistics deviation (or logit). While logits can range from negative to positive infinity, a much more restricted range is usually delineated, depending on the precision required at the tails of the curve.

While the curve in Figure 4-2 shows that students of ability \( \theta = b = 0 \) have a 50% probability of solving the item, the Rasch model does not require that the difficulty parameter be fixed. In fact if students of ability level 2 solve this problem 73% of the time and students of ability level -1 solve this problem 12% of the time, the item curve might look like the one in Figure 4-3. The difficulty of such a problem could be calculated to be 1 (i.e. \( b = 1 \)). In this case students of ability 1 (i.e. \( \theta = 1 \)) solve the item about 50% of the time.
Figure 4-3. This curve is identical to the one shown in Figure 4-2, however, in this case students of the same ability as those in Figure 4-2 are less likely to answer the item correctly. The value of the parameter \( b \) can be computed to be 1.0. Therefore, the problem has difficulty of 1 and the curve is shifted to the right one logit unit.

Given that one has an estimate of student ability \( (\theta) \), which is generally computed using student performance on the overall test, and one knows the probability that these students solve the item of interest, one can plot the characteristic item curve like those shown in Figures 4-2 and 4-3 and determine an item’s difficulty. It should also be obvious, given one accepts the assumptions of IRT, that even if the test pool is composed of students with little diversity in ability (e.g. \(-3 < \theta < 0\)), the item curve will be similar to the one shown above within that limited range.

Once a test maker determines the item characteristic curves for all the items on a test, it is argued that s/he can use these difficulty levels and an individual student’s score on each item to measure the student’s ability. Because difficulty is included, the test forms each student completed need not be identical in order to determine a “universal” score. Before accepting such a model, however, one must realize that this methodology is based on certain assumptions. First, if characteristic item curves do not fit the logistic model (e.g. students of “lower” ability answer correctly more often than their peers of “greater” ability), these items must be abandoned or “fixed” in order to use the IRT model. Second, the model assumes a student’s ability is constant in the sense that it is a value in a particular context, that there are no “carry-over” effects from other items, and that a student received no instruction between calculations of the item characteristic curve. The model assumes that a single “latent” ability is being measured. This is similar in a way to Spearman’s g although in a much more limited sense. Third, it assumes that the probability that a student of ability \( \theta \) will solve a problem is constant. Given the variability of human response to the same stimuli over time, this could be an untenable assumption. A way around this “educational conundrum” will be explored in Chapter 5. Fourth, because of the model used, items must be scoreable as right or wrong. While IMMEX problems can be scored in that manner, the power of IMMEX comes not
from determining a student got something right or wrong, but how they arrived at that answer. There is, at present no IRT model which allows for "partially" correct answers. Finally, a student ability has to be estimated in order to calculate that student's ability. This is somewhat of a "Catch-22". While there are iterative or statistical methods to do so, ultimately student performances are made to fit a model that has been defined a priori. Although this particular "limitation" will be removed in the next chapter, in order to use the IRT model, such a limitation must be acknowledged and accepted.

While Hashway observes that research has largely verified the validity of these assumptions (Hashway 1998), most of them become inconsequential in the present research in large part because the time during which students used each IMMEX™ problem was a relatively short one week exposure period. Furthermore, in every case student responses closely matched the logistic regression model which forms the basis of Item Response Theory. Finally, while IRT bases outcome solely on whether a student response was right or wrong, the use of IRT in this study is intended only as a metric of content equivalency between instructional and assessment items, not to evaluate learning. As Bloom suggests, item statistics can be a starting point in identifying ambiguous items. (Bloom, Madaus et al. 1981) Therefore, even with its limitations, Item Response Theory allows us to examine two relationships which are important to our consideration of content validity. First we can look at the relationship between items of similar type and overall student ability. Second, this tool gives us the opportunity to examine how similar items correlate with one another.

4.3.2 Student performance as content validity measure

Demonstrating content validity requires that items on a test or assessment be related to the curriculum or instructional area being evaluated. The test items should be drawn from what was taught or from general student experience not from some other area. The assumption of Item Response Theory that students with greater ability are more likely to answer an item correctly than students with less ability seems useful here. While such an assumption seems plausible for students who have had some experience with the content an item tests, given the discussion in Chapter 2, it seems untenable for students who have not been exposed to the content tested. The literature suggests that when curricular content and test content do not match, students, in general, do not score well on tests. Therefore, I conclude that if an item was not part of the curriculum, the distribution of probabilities of correct student answers when plotted against student ability would be much more random than the expected logistic function. Consequently, the correlation between overall student ability and student ability on the individual item of concern should be low and perhaps even negative. As Hashway suggests, while empirical analysis alone cannot verify the validity of how well an assessment item matches a general content area, it is part of such a verification. (Hashway 1998) Along with an analytical evaluation, this can form the basis of an argument for or against content validity.

In Hazmat, over 90% of the students are able to correctly identify the cation in the unknown compound. When looking at the student performances and answers, one quickly concludes that the students not only had the knowledge but could apply that
knowledge to determine this half of the unknown compound. The second half of the unknown (the anion), however, is harder for the students to identify. The students' attempts to solve this part of the problem lead to great diversity in problem-solving strategies and will be discussed in detail in the next chapter. The question at present is, did solving for each of these anions require knowledge that was presented within the curriculum or that students had prior experience with? In this case, an example of prior experience might be remembering that vinegar (an acid) added to baking soda (a carbonate) produced bubbles of gas. To help answer this question in a quantitative way, Paek developed models of ability on each anion for each of Ms. Lindsay's students. She then compared these models with a student's overall ability on Hazmat. (Paek 2000)

The student performances on Hazmat, in fact, suggest that the clones are all measuring from the same content area. When student performances are analyzed in terms of the anions tested by various clones in the Hazmat IMMEX™ problem, student ability on each correlates significantly with overall student ability suggesting more than just random chance accounts for such ability. Had any of the anions tested come from outside the curriculum or prior student experience, one would expect a more random association. Figure 4-4 lists these correlations.

<table>
<thead>
<tr>
<th>Ability to solve anion type:</th>
<th>Correlation with overall ability ($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chloride</td>
<td>.841</td>
</tr>
<tr>
<td>Carbonate</td>
<td>.618</td>
</tr>
<tr>
<td>Hydroxide</td>
<td>.763</td>
</tr>
<tr>
<td>Nitrate</td>
<td>.831</td>
</tr>
<tr>
<td>Sulfate</td>
<td>.747</td>
</tr>
</tbody>
</table>

**Figure 4-4.** The correlation between item types (classified by anion) and overall student ability is shown in this figure. Students with high overall ability should perform well on all the various items if the items were, in fact, part of the curriculum or prior student experience. A low correlation coefficient could suggest that even students who perform well overall might not have the knowledge or experience necessary to accomplish the task. This would suggest that the item might have been drawn from content that was not part of the curriculum. All coefficients are significant at $p < .001$.

However, the demonstration of content validity requires not just that a test or assessment be relevant to the curricular or instructional area being tested, but that it sample, and therefore represent, the entire content. IRT analysis can also provide some idea of how widely the assessment samples. In this case, instead of determining the correlation between problem dimension and overall ability, we examine the correlation between the different dimensions. While high correlations in the previous case indicated that the dimension was within a curricular area, here low correlations demonstrate that the dimensions do not require the same ability to solve the posed problem and so are not assessing the exact same aspects of the content area. Figure 4-5 lists the various inter-correlations between dimensions in the Hazmat performances from Ms. Lindsay's class.
<table>
<thead>
<tr>
<th></th>
<th>Chloride</th>
<th>Carbonate</th>
<th>Hydroxide</th>
<th>Nitrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbonate</td>
<td>.447</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydroxide</td>
<td>.239</td>
<td>.065</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrate</td>
<td>.517</td>
<td>.196</td>
<td>.381</td>
<td></td>
</tr>
<tr>
<td>Sulfate</td>
<td>.096</td>
<td>-.176</td>
<td>.678</td>
<td>.508</td>
</tr>
</tbody>
</table>

**Figure 4-5.** The inter-correlation between item types is shown in this figure. In general student ability (θ) from an IRT analysis on one item does not correlate with student ability (θ) on another item. A high correlation coefficient could suggest that different items were not measuring significantly different student abilities and therefore might not be sampling from different areas of the instructional domain.

Three inter-correlation coefficients in Figure 4-5 above could identify dimensions that are measuring the same student ability and, therefore, might be sampling from the same area of the curriculum. The largest correlation exists between student abilities to solve for the Sulfate and Hydroxide anions. As Hashway suggests, this correlation is not a definite indication that only one of the two dimensions should be retained (Hashway 1998), but it might point out an overlap in the curriculum. In fact, the various methods students use to identify the two types of clones are very different. For example, while a pH test would conclusively indicate the presence of the hydroxide ion, it would give no indication that the sulfate ion was present. Similarly, precipitation tests would also produce very different answers. In addition, Paek’s results also suggest that students of all abilities find hydroxides relatively easy to solve, while all but the most able students have difficulty solving for the sulfate ion. These differences, however, are not large enough to be statistically significant. As such, while the student performance on each dimension would indicate they are measuring similar abilities, an analytical examination suggests that the dimensions, in fact, require different abilities to solve. The correlation between Nitrate and Sulfate, and Nitrate and Chloride are not as easily dismissed. In fact, making the distinction between these dimensions is a subtle activity. Consequently, the large correlation between these pairs of dimensions could result from the fact that very similar precipitation tests are required to make the distinction. See Appendix 3 for the “solubility rules.” However, the very low correlation between Chloride and Sulfate casts some doubt on such an interpretation. Furthermore, even if the required series of tests are similar, the tests students would use to distinguish each of the dimensions lead to different conclusions, even though they might consistently follow one another in a typical student strategy. In fact, all three dimensions must be retained to fully sample the differences, albeit subtle differences, in the qualitative chemistry curriculum.

While the correlation between student performance on Hazmat and a student’s reading proficiency will be discussed in much greater detail in Chapter 6, such a correlation has important implications for content validity as well. Every written test that Ms. Lindsay’s students took during the post-treatment period of this study is significantly correlated with their reading results on the SAT-9. As indicated in the previous chapter, the correlation is often stronger between these tests and reading than between these tests and science and math measures. In other words, students who demonstrate greater reading proficiency also do better on tests in general. Such a correlation suggests that reading ability might be masking the true inferences we should be making about a student’s
ability in qualitative chemistry. Unlike these tests, a students ability on IMMEX™ problems is not correlated with SAT-9 reading score even though it is correlated with math and science grades. The obvious conclusion this seems to suggest is that, while content is important to solving IMMEX™-type problems, reading ability is not as critical and, therefore, does not confound the inferences we make about a students ability in this domain.

Before concluding this chapter, it should be noted that because the average student only accomplished 20% of the possible Hazmat clones, every student did not practice with cations and anions that later appeared in the assessment. Although one may argue that this was the choice of each student, it must also be acknowledged that while some students might merely have had to recall a practice performance, others would have had to create an entirely new performance on the assessment, which is presumably a harder task. Raizen's research suggests that student must also develop “strategic” knowledge to learn where and how to apply procedures, check progress, and obtain additional resources. (Raizen 1990) While IRT corrects for this somewhat by calculating a “difficulty” measure for all items, it cannot correct for these anomalies of “prior experience.” Although this lack of experience seemed inconsequential in this study, it could conceivably contribute to content validity issues in future applications. One way to correct for such irregularities would be to randomly assign IMMEX™ clones in a manner that ensured all cations and all anions were assigned before the student received clones that duplicated either type of ion. In addition, the teacher would have to require that students completed a number of clones sufficient to ensure the students experienced all ions at least once. Although not apparent in the present study, in other situations two practice Hazmat problems might prove insufficient to accomplish this coverage.

4.4 Conclusions
Because the current large scale assessments most often used in the United States at present do not match the instructional curriculum most students are exposed to, the data they generate can lead to inaccurate assessments of student achievement, school effectiveness or other measures of educational excellence. In order to accurately measure how well students learn, understand, and apply conceptual knowledge, one must ensure that tests measure what was taught and can adapt to the way in which it was taught.

Unfortunately, the very nature of pencil and paper tests limits their adaptability. Consequently, if the content of pencil and paper tests is designed to meet one learning standard or the curricular content of one classroom, it is unlikely to adequately represent the content of a different standard or classroom. The greater the difference between standards or classrooms, the more likely the content disparity between learning and assessment will be. Because research suggests that prior experience is an important consideration in what students can learn or achieve, a test created for one group of students, might not accurately measure another group of students. In addition, how the students were actually instructed needs to be considered when designing assessment. Research cited above, and in the last chapter, suggests that the learning and application of content cannot be “decontextualized.” One must have the opportunity to learn and
explore the application of knowledge, as well as the knowledge itself, in a way s)he will
be expected to apply it.

These considerations suggest that the most accurate evaluation of the match between
instructional curriculum and assessment content can be made at a local level. Since
teachers and students are the people who most intimately experience a given curriculum,
it is logical that they are the best evaluators of how well an assessment matches the
curricular and instructional content that was experienced. While such a paradigm of
content evaluation does not preclude “subject matter” experts from the evaluation
process, it does suggest that the process become more bottom-up than top-down. In
essence, a test is developed at the local level and “bubbles up” to higher levels such as
school, district, or state, etc. Unfortunately, written tests don’t always lend themselves to
modification. One reason for this limitation is that these tests are often designed against a
particular student model or expected response. This is especially true for multiple-choice
or other limited response tests. In addition, many of these tests must remain secret so that
they can be used to infer ability across an entire domain while sampling from a very
small number of topics. This further limits widespread evaluation of the match between
what is tested and what was taught. Therefore a dichotomy exists: we must use a single
large-scale test that evaluates all test-takers in the same way regardless of the curriculum
they were exposed to and be satisfied with inaccurate inferences about student
achievement; or we must evaluate students differently and be satisfied with inaccurate
inferences because we can not compare evaluation results. These were the quandaries
suggested in this and the previous chapter.

Depending on how it is used as an assessment tool, modern computer technology offers
U.S. educators and policy makers an alternative to this dichotomy. If educators or policy-
makers decide to use technology merely to replace pencil and paper tests, the research
suggests little improvement in our inferences about student learning will result.
Conversely, if educators and policy makers take advantage of the adaptive nature of
technology, more accurate inferences are likely. This chapter details how one such
technology, IMMEXTM, positively affects education by ensuring a closer match between
the instructional curriculum and assessment.

IMMEXTM allows educators to build assessments more closely tailored to the
instructional curriculum delivered in three ways. First, the teacher can author or tailor a
problem to match the needs of his or her curriculum and students. As was demonstrated
in the present research, Ms. Lindsay reduced the domain of the problem to match the
content of the classes she taught. Another teacher could easily increase the content
covered by an IMMEXTM problem just by adding information to the problem space.
Second, because the problems are available in the public domain via the World Wide
Web, they are accessible by “subject matter experts,” policy-makers, educators, students
and parents. Consequently, the match between assessment and curriculum or standards
content can be reviewed at any level. Most significantly, because the assessment is no
longer secret, students are able to know how they will be assessed and to challenge any
content they feel is inappropriate before, during or after being assessed. Finally,
IMMEXTM allows educators to assess how well students can apply the content they were
exposed to, in the way the students were exposed to it. Because IMMEX™ records all the inquiries a student makes between the presentation and solution of a problem, educators can truly observe how that student attempts to solve the problem. Apparently the resulting data is not confounded by the student’s ability to read English. In addition, the technology allows the form of student assessment to match the form of instruction and student learning. While this is especially important in “hands-on” domains like the sciences, it is equally important in any domain where there is an interest in the ability of students to apply their learning in the real world (e.g. music, math, history, languages, etc.).

As the results in this chapter suggest, in light of current standards, IMMEX™ demonstrated a very high level of content validity as it was employed during this study. Not only did Ms. Lindsay create the Hazmat problem, she was also able to modify the problem to accommodate the specific group of students she had in her classes and the content to which her students had been exposed at the time IMMEX™ was introduced into the classroom. She also indicated IMMEX™ allowed her to assess aspects of learning she had, heretofore, been unable to evaluate.

Students, too, acknowledged that IMMEX™ aligned well with the chemistry curriculum they had experienced. This “acknowledgement” came both from survey results and from the evaluation of student performance. As reported above, student performance suggested a strong correlation between overall student ability and ability to solve particular Hazmat clones. This significant relationship lends strong support to the belief that the students had been exposed to the instructional content assessed in each of the clones. Furthermore, the general lack of correlation between student ability on different clone types suggests that the various clones measured different aspects of the instructional curriculum.

One more important benefit of this type of technology is important to mention. As alluded to earlier, it is virtually impossible to determine the validity of a local educator’s assessments at present. However, these types of assessments are the most ubiquitous of evaluations in present day American education. IMMEX™, coupled with the technological and statistical tools presented in this chapter offer a nascent methodology to measure the validity of these tests.

This chapter demonstrated how content validity might be improved. As will be shown in the next chapter, student performance on IMMEX™ problems like Hazmat also suggest similar improvements when we want to infer near- or long-term student performance.
Chapter 5. Determining what students can actually do

In the preceding chapter, we looked at the relationship between the material a student was exposed to in a curriculum and the material covered on a test. As such we concentrated on what was being tested. In this chapter, we turn our attention to how student learning will be evaluated, and what criterion will be used to judge student ability and achievement. According to Guion, what is appraised in a criterion oriented study is the validity of the hypothesis of a relationship between the test and a criterion measure. (Messick 1989) The criterion measure may be something as simple as performance on the next test, or the consistency of student performance over time. This performance can either be measured dichotomously or in a more holistic fashion. Resnick observes that standardized multiple choice tests or other correct/incorrect measures fail badly when judged against the criterion of assessing and promoting a thinking curriculum. (Resnick 1987) While easy to evaluate, standardized, multiple choice tests really only tell us whether students can recall a certain piece of information or pick that piece of information out of a list of pieces of information. The “thinking curriculum”, on the other hand, requires that students actually apply learned content to real life or authentic problems in whatever domain is being assessed. As its name suggests, such a curriculum requires students to think about what they have learned, not just memorize facts. In fact, Maker cites a great deal of research to suggest that when assessment occurs under the actual conditions a performance will be made or evaluated, predictions about future performances are much better. She also suggests that these performances should be made using various problem types and in various settings. (Maker 1994) This is similar to the multiple modes of assessment advocated in previous chapters.

I have also advocated the need for greater public accessibility in testing. This accessibility should cover not only the content of the test, but also the criterion against which students are tested. Wiggins observes that the indication a student is beginning to make real progress correlates well with the accuracy of the student’s self-assessment. Accurate self-assessment requires students understand the criteria against which they will be judged. In all these cases then, we need some standard against which to measure student performance.

5.1 The need for standards (criteria)

In order to effectively evaluate student performances, we must agree on what the student should be able to know and do. We must have a standard. At present, large-scale, multiple choice tests require students pick the correct answer from a list of possible answers. Tests composed of more open-ended questions and actual performance assessments, on the other hand, usually have a more complex gauge against which to measure success. This is often referred to as an evaluation rubric. Even a classroom teacher will have to articulate some standard against which to measure student performance. Students, to evaluate their own progress, need to have some idea of the models and standards against which to appraise the value of their performances. These student standards are probably heavily influenced by the standards set for them by their teachers. The evidence in this study supports such an observation. Students
overwhelmingly agree (86%) that their teacher's evaluation of their qualitative chemistry ability is accurate, and the correlation between teacher and student self evaluations is significant. While the standard or criteria of success could be as simple as specifying the correct solution to a posed problem, such dichotomous criteria seem insufficient to evaluate what the current standards suggest is important. To know if students "use their minds well" and can actually apply what they have learned, we need to know how the student arrived at an answer, whether s/he will be able to reliably solve similar problems again, or if there are flaws in or a lack of student ability that may undermine student performance.

Wiggins argues that effective standards are not merely arbitrary inventions of an individual. (Wiggins 1993) All stakeholders must both understand and accept them as important. The standards must also be realistic. If the standard requires that which is beyond a system to achieve, they are not effective. Many of the states have experienced the need to publicly defend their standards both in the courts and in the media as they began denying high school diplomas to students who did not meet their high school graduation criteria in the late 1990's. In fact Wiggins argues that effective standards are not only required for accountability, but more importantly that they are required so that those being assessed might know how to improve their knowledge or skills. He demonstrates the futility of an uninformative standard by describing a classroom he once observed. An English teacher would write "vague" on papers and return them to the students to redo. At the end of the year, one of the students asked the teacher what "vagoo" meant. Obviously the student not only failed to understand the feedback, but also had little understanding of the standard against which the teacher was evaluating his work and so could do little to improve. Without accurate, understandable, and achievable standards not only will our inferences be ineffective and overly broad, but the path a student or educator should take to improve student performance will not be clear. Furthermore, while we may be able to use such standards to conclude a group of students cannot perform at some required level, we will probably be unable to determine the degree of ability each student actually possesses. In order to begin making their own self-evaluations, students must learn to both produce and recognize quality work. This will only happen, according to Wiggins, if standards are so transparent as to make student self-assessment accurate. (Wiggins 1993) Therefore, in any local or state assessment, evaluation should be objective - not subjective, should reliably produce consistent results, and, most importantly, should be understandable to the student and other stakeholders. Moreover, to truly affect education, it must also be public. What we need are consistent criteria against which all stakeholders can compare student performance.

5.2 National science standards
In the United States, the federal constitution implicitly vests the responsibility for education in the states. As such, the federal government has no authority to create binding standards. While assessment standards have been created for the NAEP, the states have effectively resisted anything beyond this. In an effort to fill the "standards void" at the national level, the National Research Council has established an objective, reliable, and
public national science standard. (NRC 1996) Among other things, the grade 9 - 12 Science Education Standards they produced indicate that students should:

- Demonstrate appropriate procedures, a knowledge base and conceptual understanding of scientific investigations.
- Use evidence, apply logic and construct arguments for scientific investigations.
- Formulate and revise models using logic and evidence.
- Use scientific criteria to find preferred explanations.

In many ways, these standards parallel the Goals 2000 standards in that students will “be able to use their minds well” in order to pragmatically apply science in their lives.

5.3 California state science standards
The State of California standards are quite different. Perhaps this is not surprising after the previous discussion of how the inferences made from state and NEAP test results differ. Over the course of twelve pages, the California standards emphasize what students should know. Only in the final page of the standards are students required to move beyond what Bloom calls the most rudimentary level of learning. The last page of the California State standards indicates that students should understand that scientific progress is made by developing meaningful questions and by conducting careful investigations to try to answer those questions. They cite, as the basis for evaluating a student’s understanding of the content addressed in the curriculum, an ability to formulate questions, develop explanations, and an ability to analyze problems that require students to combine and apply concepts across multiple domains of learning including science. Except for this final page of the California standards, however, there is not a lot of emphasis on what students should be able to construct, demonstrate, formulate, and analyze like that found in the “national” standards. Instead, the state standards focus on what the students should know or be able to recall. As we have seen, student recall is heavily tested in the state’s student achievement test (the SAT-9) even though test and curricular content may not necessarily match. Unfortunately, this test is also largely unable to assess student performance abilities.

Student knowledge, however, is not the same as understanding. Recall is not the same as the ability to apply what is learned. Curiously however, the California standards also state that “the guidance in the science content standards for California’s K-12 schools is not binding on local education agencies or other entities.” Basically the state standards are a guideline, not a firm standard against which California teachers can measure student progress. Rather it is a voluntary metric that school systems or teachers are free to accept or reject. Therefore, not all California teachers may be using the same metric to evaluate their students. Consequently, assessment criteria can vary not only between districts and schools, but from classroom to classroom. Nevertheless, for our purposes in this research, it is important to observe that while the weighting on knowledge and performance might be different, both the state and national science standards recognize the need for educators to measure both knowledge and performance in evaluating student ability. Consequently, one would expect students in California to be assessed on their command
of content and their ability to apply that content. Unfortunately, the current state test of academic readiness was not designed to test the latter.

### 5.4 Reconciling the Standards

The tension reflected between the state and national standards or the state standards and the state assessment in California is not new. Bloom cites these two needs of evaluation as the basis of the two predominate models in the modern instructional objective movement: task analysis and the Tyler approach. (Bloom, Madaus et al. 1981) The former model originated in training personnel for World War II and the Korean conflict. The goal of task analysis is first to identify a desirable or necessary skill, and then isolate or decompose the necessary behaviors that must be learned to successfully perform that skill. These behaviors can then be both taught and assessed. In many cases the decomposed behaviors can be assessed with pencil and paper tests. This idea forms the basis of what constitutes a large part of educational assessment today. Depending on the measurement criteria, this type of teaching and assessment may be entirely warranted. As Bloom notes, “in training, unlike general education, learning a sequence of steps by rote will ensure that the trainee acquires the desired skill.” (Bloom, Madaus et al. 1981)

However, as has been noted repeatedly in previous chapters, this may not ensure that students can actually perform as well on dissimilar tasks in the short term or long term future, especially if these future performances do not involve pencil and paper, multiple choice tasks. Bloom continues by noting that “skills form a basic component of the educative process,” and that there are numerous “skills in most areas of the curriculum that we want all students to acquire.” In fact, much of the criticism of the educational system today seems targeted at the fact that students are graduating without the “basic skills” necessary to function effectively in the larger society. However, such contextual dependency may be antithetical to the broader goals of education. While “training” may be appropriate for some learning, Bloom concludes that mere rote learning will still not provide adequate education. (Bloom, Madaus et al. 1981)

Tyler, on the other hand, saw education as the production of behavioral changes in the learner, and assessment as an evaluation of these behavioral changes in relation to specific desired behaviors. For Tyler, statements of educational objectives served as the basis of both curriculum development as well as guides for the creation of assessments. In this model, objectives are broken down into a content and a behavioral component. Since it is unlikely a behavior in each content area can be evaluated on a single test, summative tests usually merely sample from the full range of objectives. Assigning relative values to various behaviors and sampling from the content domain ensures content validity, while looking at how the student actually solved a problem behaved ensures criterion validity.

The two approaches agree in some common areas: students must be able to routinely “do” something after instruction that they could not necessarily do before; instruction must be accountable for success; and objectives must be defined in terms of desired actions, products or both. What they disagree on is the value each places on rote
memorization of knowledge or the logical application of that knowledge, and the student’s ability to adapt to unforeseen or unexpected situations.

As both the national and state standards stress the need for students to apply concepts from their course of study in order to develop an understanding of science, it seems only logical that our assessments of students’ science understanding should allow us to evaluate both knowledge and performance. As previously argued, the evaluation of students’ prior experiences is only valid if the students all have the prior experience over which they will be tested and the format of the test matches the format of their experience. For example, if they are taught using hands-on learning, they should be assessed using that format. Furthermore, as discussed above, the most effective assessment of future performance is an assessment of ability done in a manner as similar to the actual performance as possible. Multiple choice and other pencil and paper tests, don’t accomplish this well. Obviously then, the best we can do with many multiple-choice tests is to measure student performance against how well they are able to recall facts in certain settings. Often this devolves into whether the student has answered correctly or incorrectly. With only a single degree of freedom, we are only able to get a very limited sense of a student’s ability to apply knowledge. While Minstrell and others have reported varying degrees of success with attempting to infer different levels of student ability based on answers to multiple choice questions (Minstrell 2000), we generally have a difficult time inferring any sort of student ability in such a decontextualized setting. For example, what Ms. Lindsay really wanted to know in her classroom was whether the students were able to “logically” solve a problem, and she felt unable to make accurate inferences of this using pencil and paper tests.

Webster’s dictionary defines logical as “correct reasoning” or “that which is to be expected because of what has gone before.” Wiggins suggests that this logic should be demonstrated in real world settings. (Wiggins 1993) Consequently, to assess the logic of a student’s mental model, we need to assess the logic of their strategy for solving a problem. We need to look at student behavior as described by the Tyler model. This is exactly what Ms. Lindsay wanted to investigate in her students. Unlike the pencil and paper tests she gave her students where she could see, at best, how well they could regurgitate content or apply some memorized algorithm, what she really wanted to see was whether the students had developed a strategy to deduce the identity of an unknown compound. In other words, she wanted to see if the student could use chemistry content and prior experience to formulate, construct and analyze evidence to reach a reasonable conclusion based on evidence. The standardized tests at her disposal did not allow her to do so. Even if she could observe student performance, the difficulty of how to holistically evaluate these strategies arises, especially if the standards do not specify particular performance levels. However, if we only consider whether a student was correct in proposing a solution to a problem, our inferences about how well students have achieved Ms. Lindsay’s criteria may be invalid. This was discussed previously as a major flaw with many current assessments.

Hurst, et al. suggests that technology may help overcome the limitations of traditional studies of students’ mental models and behaviors. (Hurst, Casillas e: al. 1997) If
technology is appropriately integrated into the curriculum for both practice and assessment, it can allow us to eavesdrop on student solutions and every action the student took to reach that conclusion. Consequently, we can develop an intricate picture of how one or more students actually solve problems and we can catalog these strategies in order to gain insights into how the student thinks about a problem. In other words, we can watch how a student develops a solution and not just evaluate the resulting solution.

5.5 Using General Test Theory to evaluate student ability

Because IMMEX™ technology produces large amounts of disaggregated data, we can use General Test Theory and the Rasch model explained in the last chapter to analyze the Hazmat data generated by Ms. Lindsay's students. Since the model calculates a degree of difficulty for each Hazmat clone, we can even use IMMEX™ to rank order students even though most students did not accomplish identical Hazmat problems. In fact, the correlation between Ms. Lindsay’s February ranking of the students and their individual abilities (θ) calculated based on their Hazmat performance, shows a significant negative correlation (r = -0.252, p = .003). Student ability thetas also show a strong correlation with student self-evaluation from the second student survey (r = 0.257, p < .001). In addition, these thetas show significant correlation with a student’s final exam grade (r = 0.239, p = .006) and final grade in the course (r = 0.292, p = .001). As mentioned previously, research suggests that, unlike other normative testing, these evaluations are event free because evaluations of student ability are independent of the group used to calibrate the items. They are also sample free because evaluations of student ability are independent of the actual items a student tested on.

The present research and other research previously cited suggests that Item Response theory can indeed provide some significant insights into a student’s ability, and these ability measures may be predictive of concurrent and future performance. Surprisingly, the ability measures estimated for each student based on his or her Hazmat performance is not correlated with similar ability estimates based on his or her performance in Desperately Seeking Solution. The fact that the students had only half the performances in Desperately Seeking Solution as they had in Hazmat could account for some of this. However, these estimates all require some previously defined model of performance (in this case the one parameter logistic model) that is unresponsive to changes in classroom context or student cognitive growth over time. This could also account for the lack of correlation. The expected correlations between IRT ability measures suggesting possible gender, SES, and ethnicity bias were not evident in this research.

5.6 Using percent to evaluate ability

To avoid the need to fit student performance to a static, a priori model, some have suggested using simple percentages to evaluate actual student performance in much the same way as most pencil and paper tests are currently evaluated. When using IMMEX™ technology, a percentage score is easily calculated by dividing the number of clones a

---

5 The correlation is negative because in Ms. Lindsay’s evaluation scheme one is the best, whereas in IRT logits run from -3 (worst) to +3 (best).
student answered correctly by the number of clones a student attempted. On cursory review, this measure seems to be a good predictor of concurrent and future student performance. The correlation between the percentage of all Hazmat clones a student answered correctly and a student's semester grades are significant ($r = .251$, $p = .004$ with first semester, and $r = .304$, $p < .001$ with second semester). The percentage of Hazmat clones a student correctly solved is also correlated with GPA ($r = .211$, $p = .015$) and with a student's self-evaluation ($r = .212$, $p = .019$). In addition this variable is uncorrelated with gender, SES, and ethnicity suggesting IMMEX™ performance measures may not suffer the bias problems raised for other tests. Furthermore, the percentage score is also uncorrelated with a student's SAT-9 reading score and is, therefore, probably not confounded by student literacy issues. Finally, these correlations suggest that IMMEX™ performance is dependent, in part, on how well a student has mastered the content of the course and other school subjects as measured by pencil and paper tests. Unfortunately, the percentage of Hazmat clones a student solved correctly in practice is uncorrelated with the percentage of clones $s/he solved in assessment. Furthermore, the overall percentage of Hazmat clones a student solved does not predict the percentage of Desperately Seeking Solution clones a student will solve two months later, implying the criterion validity of such a measurement is low.

5.7 Using Performance Index (PI) to evaluate student ability

One of the problems with using a simple percentage calculation to measure student achievement and ability stems from the fact that the percentage measure does not differentiate between students who attempted many clones, and those who attempted only one or two. For example, if a student attempted one clone and solved it correctly, s/he would have a percentage score of 100%. Similarly, if another student attempted and solved ten clones, their performance would also equate to 100%. As with any statistic, the more data one has with which to make inferences, the more accurate the inferences will be. In this case, a conclusion that the latter student understands qualitative chemistry is probably a more accurate inference than that same conclusion about the former student. For this reason, among others, Stevens has suggested weighting the simple percentage by a factor of the number of clones a student answered correctly. (Stevens 2000) He calls this variable the "performance index." The performance index is computed by squaring the number of clones a student answered correctly and dividing that product by the number of clones a student attempted. In effect, this metric has the effect of rewarding a student for attempting and correctly answering more clones. It may also provide a more accurate representation of how well a student understands the topic addressed by a particular IMMEX™ problem.

Like a straight percentage measure, the performance index (PI) of the students using Hazmat in this study correlates significantly with both semester grades ($r = .311$, $p < .001$ with first semester grade; $r = .353$, $p < .001$ with second semester grade), and with student self-assessment ($r = .254$, $p = .005$). The PI is also correlated with overall GPA ($r = .253$, $p = .003$), and does not show any correlation with gender, SES or ethnicity. Unexpectedly, the PI metric shows only marginal correlation with student performance on the qualitative chemistry written test ($r = .158$, $p = .071$). As with percent, however,
the PI does not predict future student performance as measured by the identical metric on future Hazmat clones or on the Desperately Seeking Solution problem. Perhaps these shortcomings are inherent in the metrics themselves.

Simple percentage, PI, and the Rasch model in IRT all suffer from a significant limitation. Each of these measures relies only on whether a student arrived at the correct answer to a clone. None of these metrics delves into how a student went about arriving at an answer. They do not fully assess what Ms. Lindsay and the standards want assessed; however, this is precisely the insight the IMMEX™ technology can provide us. While IMMEX™ can collect the rudimentary data for these simpler metrics and analysis suggests a student’s content knowledge is being assessed, such simplistic metrics do not exploit the tremendous power of this technology. IMMEX™ allows us to look beyond what a student “knows” by offering insights into what Bloom terms a student’s ability to analyze, evaluate and synthesize and what Tyler termed behavior. Although the above metrics demonstrate the use of technology to collect data to support straightforward statistical analysis of student performances, they are not robust enough to account for the level of information necessary to make the more substantial inferences demanded by the standards and by Ms. Lindsay. To accomplish this requires new tools that will allow one to analyze the large amounts of data IMMEX™ generates and tease out the detail contained therein.

5.8 Using Artificial Neural Nets to evaluate student ability

When analyzing the IMMEX™ performances of students, the complexity of the task is immediately obvious. As the number of menu items in an IMMEX™ problem increases, so too does the way a student may move through the problem space. If one considers the order a student viewed different pieces of information, the number of paths through the problem space increases exponentially with each additional information item. Even if one only considers what information a student reviewed in arriving at an answer, the number of strategies a student could use to solve the problem is a multiple of the number of information items. Such a large number of possibilities make it almost impossible for humans to consistently find a pattern in student performances.

Increasingly, the social sciences such as education seek to explain outcomes based on the simultaneous interactions of numerous variables. We want to make meaning of the events around us. In the present research, we want to make valid inferences about student learning and understanding. In order to do so, we must attach accurate meaning to the student performances at our disposal. As Messick so aptly notes, “whenever an attribute, event or relationship is interpreted or conceptualized, it is judged as belonging to some broader category to which value already attaches.” (Messick 1989) The previously cited research in cognitive science suggests human beings also make meaning of the world in just this way. This is just as true for the activities of the people around us as it is for other events in the world. In fact Fisher states that “people’s activities form coherent patterns, and analysis of those patterns is essential for understanding their activities.” He goes on to observe that a systematic analysis of the structure and organization of these activities allows researchers to find order in the seeming chaos. (Fischer and Bicell 1997)
This is not the same as comparing activities to some predetermined model. Lesh and Kelly suggest this is precisely the shortcoming inherent in comparing novice performances to experts (Lesh and Kelly 1996) and Rogoff suggests similar problems when evaluating student performances without considering the context of those performances. (Rogoff 1998) Furthermore, the work of Winnograd suggests we will never be able to build a static model of all the ways students will use to solve problems posed to them. (Winograd and Flores 1986) This was one of the problems with the IRT model cited in the previous section. Rather than comparing these performances to some a priori model then, our needs require a more dynamic and adaptable "model".

Various authors have suggested numerous ways to discern patterns or similarities in data sets such as the student performance data generated by IMMEX™. For example, Afifi & Clark suggest K means statistical clustering (Afifi and Clark 1996); Bobick describes using modal properties (Bobick 1987); and Principie, et al. suggest artificial neural networks (Principe, Euliano et al. 2000) to discover patterns in data. Because K means clustering relies on a defined statistical model and Bobick’s modal properties analysis has not been extensively developed, I have investigated the use of supervised and unsupervised artificial neural networks to discern similarities in student performances. Artificial neural networks provide an adaptable pattern recognition system able to find inherent patterns in data. Furthermore, Marshall and Zohar suggest that “neural networks offer a model for pattern recognition, or learning by association, that has in turn become a widely accepted model for some types of human perception and learning.” (Marshall and Zohar 1997) Moreover, the patterns discerned by neural network analysis are not subject to the hidden biases present in human evaluation. In fact, since one can access the numerical weighting such a network assigns each input, one can determine the input’s contribution to each observable pattern. No matter what technique is used, however, “the major decision as to whether the cluster analysis has been successful should depend on whether the results makes intuitive sense.” (Afifi and Clark 1996)

In particular, Kohonen artificial neural networks have demonstrated an ability to meaningfully cluster performances in data sets like those produced by IMMEX™. Like other Artificial Neural Networks (ANN) trained with unsupervised learning, Kohonen nets “find” clusters of data points in a large data set by moving a pre-determined number of “markers” to the mathematical center of the data. The neural net locates this center by moving the markers until the distance between them and the data points in the data cluster are minimized. Unlike supervised learning, the learning of unsupervised networks proceeds without comparing the output of the network to some external measure of what “should be.” Instead the unsupervised ANN is designed to minimize a built-in distance metric. As is often the case, the ANN used in this research was designed to minimize Euclidean distance between the markers and data points. As each new data point is introduced to the net, the ANN calculates the distance between each of its markers and the new data point. The net then minimizes the distance between the marker(s) and the data point by moving the marker(s) toward that point.
In addition to deciding which distance metric to use, the designer of ANN must also set four or more other parameters to control network sensitivity to new data and how fast the net will learn. In this research, the network parameters set prior to training included learning rate, neighborhood radius, conscience, the number of unique clusters to separate data into, and the number of times (epochs) a data set would be input to a neural network during training. I will now review these concepts and then describe how the parameters for the ANN used in this research were set.

5.8.1 Learning rate
The distance each marker moves is pre-determined by the creator of the ANN and is an important consideration in the network's design. This distance is termed the learning rate of the network. Although a large learning rate means the net will learn quickly because the marker(s) quickly move toward new data points, too large a learning rate also implies the markers may over shoot the centers of data clusters. In addition, because a marker's current position was determined by previously introduced data, allowing markers to quickly move away from their present position means these markers will give new data too much weight and cause the net to quickly "forget" the effects of previous data points. On the other hand, too small a learning rate can dramatically increase training times and make the network unresponsive to new information. As a moderate alternative to the two extremes, creators of neural nets often begin with a large learning rate, thereby allowing markers some quick but gross alignment with the data, then gradually reduce the rate to allow the markers to optimize their positions.

5.8.2 Neighborhood radius
Were every marker allowed to move so as to minimize the total distance between itself and the new data point, all markers would quickly converge at the center of the entire data cluster. To force individual markers toward specific clusters of the entire data set, network designers often make the markers myopic by allowing them to see data only a limited distance from their current location. This distance is called the neighborhood radius. As with learning rate, designers must consider the value of this parameter when designing neural networks. Too large a neighborhood means the network may be unable to distinguish individual clusters of data. Too small a neighborhood may mean new data falls outside the view of all the markers and, as a result, no marker moves to include the new data in its cluster of data. As with learning rate, gradually reducing neighborhood radius over time can have significant benefits. Initially training with a large neighborhood radius allows the network to capture the entire data set including extreme outliers, and gradually reducing this radius during training forces individual markers to specialize on the specific data clusters nearest them as training proceeds.

5.8.3 Competition and Conscience
To promote further specialization, such as in the case where the neighborhoods of two or more markers overlap, a "winner-take-all" strategy is often adopted by ANN designers. In this scheme, only the marker nearest the new data point is allowed to move toward the
new data point. The effect then is that only the single winner of the distance competition is influenced by this new data, so only this marker includes the new data in its cluster of points. One drawback to this approach, however, is that if a set of data points were closely grouped or presented to the network in a certain order, a single marker may end up winning all the distance competitions. The result would be sub-optimal data clustering and noncompetitive markers that had few, if any, data points in their vicinity. These are also called dead markers”. To prevent this, the winner-take-all strategy is given “conscience” by network designers.

Conscience ensures that the more an individual marker wins, the harder it becomes for that marker to win future competitions. A small positive number, which I call it the “guilt coefficient” or gamma, penalizes the marker or makes it feel guilty each time it wins a competition. The penalty is actually the fraction of the time a marker wins a competition. Therefore if gamma = 0, the marker would suffer no penalty for winning. If, on the other hand, gamma = 1 the marker would incur a full (100%) penalty for each win and no penalty for each loss. In order to actually apply this penalty in each competition, the network uses the penalty to compute a distance bias. This bias is then subtracted from the distance measurement computed when new data is introduced to the network. The bias is computed by subtracting the penalty from the expected frequency each node should win (estimated to be 1/number of nodes) and multiplying the result by a positive constant. I call it the “bias coefficient” or beta. If beta = 0, no distance penalty would ever be assessed and the network would function without conscience. Conversely, if beta = 1 then as the penalty approached the expected frequency each node should win, the distance bias would become smaller and smaller. As such beta represents the degree which deviations from expected wins will effect clustering. When the penalty factor exceeds the expected frequency a node should win, the bias becomes increasingly negative. The bias is then subtracted from the distance measured between the marker and each new data point. Thus as the bias becomes more and more negative, the distance used to determine a “winning” marker becomes larger and the node wins less often. As with learning rate and neighborhood radius, the designer of the net builds gamma and beta values into the net at design time. Although a wide range of values should work (Principe, Euliano et al. 2000)6, if gamma is small the conscience will have little effect on markers winning too frequently. Antithetically, a large gamma can over penalize winning and prevent effective learning. Similarly, if beta is too large, the distance penalty for even slight variations from the expected winning frequency will be severe. Too small a beta allows a single marker to win more often again raising the problems conscience attempts to reconcile. Figure 5-1 details the actual range of values that were tested in designing the ANN.

5.8.4 Soft Competition
A unique strategy employed by Kohonen ANN is soft competition. Although Kohonen networks determine a winning marker as each data point is introduced to the network, unlike other schemes Kohonen nets also move the neighbors of the winning marker a

---

6 Principe, et Al. labels the bias coefficient as gamma and the guilt coefficient beta (exactly opposite the labels used in this dissertation).
small amount that is usually proportional to the distance each neighbor is from the data point.

5.8.5 Determining ANN parameters

Although unsupervised ANN require no a priori knowledge of the data, few explicit rules exist for determining the values of the network parameters just discussed. Therefore, while not absolutely required, some knowledge of the data can be helpful in choosing the various network parameters required in designing the network. The designer of Kohonen ANN must set these parameters when the network is created. However, because there are virtually no explicit rules for determining the value of such parameters, experience and testing are useful in choosing optimal values in such designs. In the present case experience suggested the number of markers to use, while actual performance on test data suggested values for the other variables.

Stevens has suggested that the complexity of the particular IMMEX™ problem should determine the number of different data clusters or unique strategies one can expect to result from numerous performances of an IMMEX™ problem. (Stevens 2000) His experience suggests that student performances in solving problems with approximately 100 different items of information can be adequately described by 25 different clusters. Performances on problems with 1000 or more menu items are probably separable into approximately 100 clusters. Given the number of menu items in Hazmat and Desperately Seeking Solution, I designed an ANN to group student performances into a maximum of 25 different clusters.

Trial and error suggested the value of the other network parameters. The ANN in this research was designed to discern patterns in data sets representing up to 90 different information items a student could choose. As such, each item of information could be considered a separate “dimension”. However, actually confirming how well a network finds data clusters in a data set of such large dimensionality is difficult because it is hard to visualize. To avoid over-fitting the network to a specific data set, I created sets of two dimensional sample data. Each set of sample data contained various patterns of clustered data. Furthermore, since these clusters were only in two dimensions, it was easy to actually see if the networks had accurately clustered the appropriate data points together. Figure 5-1 lists the parameters and the range of values used to test each parameter during these initial tests.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Low value</th>
<th>High value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood Radius</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Conscience: Gamma</td>
<td>.001</td>
<td>3</td>
</tr>
<tr>
<td>Conscience: Beta</td>
<td>.1</td>
<td>10</td>
</tr>
<tr>
<td>Number of Markers</td>
<td>1</td>
<td>25</td>
</tr>
</tbody>
</table>

**Figure 5-1.** The parameters and range of values (low to high) of each parameter tested during two dimensional training of a Kohonen artificial neural network to cluster test data.
The initial placement of the markers in the 2-D test space is a random event when training ANN. Depending on where these markers started, low fixed values of neighborhood radius produced large numbers of dead markers and the ANN seldom found all data clusters. On the other hand, high neighborhood radius values resulted in the convergence of all markers in the center of the entire data set and no identification of individual data clusters. Although a fixed neighborhood radius of 5 produced the best results, markers occasionally had difficulty locating individual data clusters. Rather, they settled midway between data clusters. This suggested a gradual reduction of the radius might be warranted.

I also tested the net with diverse learning rates. While various learning rates at the low end of the above range were slow to converge, given enough time and the random initial placement of the markers, with learning rate greater than zero all markers eventually settled on individual clusters of data. A learning rate of one produced marker placements closest to the centers of individual data clusters; however, without reducing the learning rate over time, markers were seldom truly centered within data clusters.

Of all parameters, the conscience parameters (gamma and beta) proved the most difficult to determine. Principe, et al. recommends a gamma value of .001 and a beta of 3.7 (Principe, Euliano et al. 2000) Nevertheless, these numbers made the behavior of at least one test net unreliable. Generally beta values greater than one tended to draw all the markers toward the center of data set, while lower betas forced the markers to divide up the data set more evenly. Similarly, gamma values greater than one resulted in a single greedy node winning all competitions. To promote greater marker dispersion and a more equal number of data points around each node, I lowered beta and gamma values until the net produced more optimal results on varied two dimensional data sets. The resulting values (Beta = .3 and Gamma = .01) were used in all nets designed to find clusters of IMMEX\textsuperscript{TM} performances.

Although experience suggested 25 markers would most appropriately describe the performances in the IMMEX\textsuperscript{TM} data sets we examined, I conducted tests to ensure the net would act appropriately when it had either more or fewer markers than actual clusters in the data. In all cases, the markers divided the data set appropriately with few, if any, markers defining empty data clusters.

As described above, although a large neighborhood radius ensures no data point lies outside the scope of network markers, ultimately every marker will converge at the center of the entire data set and will not divide the data into individual clusters if the radius remains large throughout the training period. The pros and cons of small neighborhood radii are also presented above. For this reason, gradually reducing the neighborhood during training is a suggested strategy. By trial and error I found reducing the neighborhood radius from 4 to 0 by a linear factor of -0.004 between epoch 10 and epoch 7000 of training produced the best results. For reasons also presented above the learning

---

7 Here again, the authors use beta for the "guilt" coefficient and gamma for the "bias" coefficient (the reverse of what I use in this dissertation).
rate was reduced as training proceeded. Here again, I used trial and error to determine how to schedule the reduction of the learning rate. Ultimately a reduction of -0.00015 from a learning rate of one to zero between epoch 10 and 7000 proved most appropriate. Changing the learning rate and neighborhood radius required training time be lengthened to ensure the markers had “time” to locate the center of each data cluster. The value of 7000 training epochs was chosen because it is large enough for the markers to converge on the data centers, but small enough to ensure the ANN would generalize to data outside the training set. Exponential and logarithmic schedulers for reducing learning rate and neighborhood radius proved less optimal than a linear reduction during my tests.

In summary I tested a square Kohonen ANN constructed with the following parameters on various 2-dimensional data clusters and determined the values in Figure 5-2 produced optimal results for all the nets tested:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of training epochs</td>
<td>7000</td>
</tr>
<tr>
<td>Number of markers (nodes)</td>
<td>25</td>
</tr>
<tr>
<td>Neighborhood Radius (Initial)</td>
<td>5</td>
</tr>
<tr>
<td>Radius reduction (per epoch)</td>
<td>-.004 (between epochs 10 and 7000)</td>
</tr>
<tr>
<td>Learning Rate (Initial)</td>
<td>1</td>
</tr>
<tr>
<td>Rate reduction (per epoch)</td>
<td>-.00015 (between epochs 10 and 7000)</td>
</tr>
<tr>
<td>Conscience (Beta)</td>
<td>.3</td>
</tr>
<tr>
<td>Conscience (Gamma)</td>
<td>.01</td>
</tr>
</tbody>
</table>

Figure 5-2. These parameters produced the Kohonen ANN most able to find clusters in various 2 dimensional data sets. The same parameters were used for Hazmat and Desperately Seeking Solution problems.

5.8.6 ANN trained for larger dimensions

The same neural nets developed for trial, two dimensional data were then used to determine the data clusters in the Hazmat and Desperately Seeking Solution problem. In each case the only network parameter I changed for different IMMEX™ problems was the number of input dimensions. This change was necessitated by the nature of each specific problem. For example, in Hazmat the student can choose approximately 23 different information items. In the Desperately Seeking Solution problem, each student can choose up to 90 different menu items. As described above, this is akin to comparing student performances across 23 or 90 dimensions simultaneously. Therefore while Desperately Seeking Solution requires 90 inputs, Hazmat requires only 23.

Inevitably, the question arises of what constitutes an appropriate data set for training an artificial neural network. Ultimately, the answer is that one should use a set of data that contains exemplars truly representative of the performances one eventually desires to cluster. Included in such deliberations must also be the eventual use of the clustering.
These questions are similar to the ones that should be asked of any assessment criteria. In training ANN, however, we must ponder the question of whether training with only certain student performances is varied enough to cluster the diversity of student attempts to solve the problem. The decision to use only successful student solution strategies to train the unsupervised net in this research was made for two reasons. First, experience suggests that students successfully solve IMMEX™ problems using a variety of strategies and that these strategies vary over time. In fact, the Hazmat data contains successful performances in which the student obviously guessed at a correct solution. In general, adding unsolved performances to the training set would not diversify the solution strategies on which to base the data clusters discovered by the net. In fact, adding unsuccessful performances to the training set would merely add a lot of disparate, infrequently used, and unsuccessful strategies to the training set. This in turn could require the net to dedicate markers to unsuccessful strategies used by small numbers of students, while simultaneously forcing the net to group successful student strategies into larger clusters of more generalized strategies. Secondly, using a wide variety of successful performances ensures that unsuccessful strategies are ultimately compared to successful ones. In this way, we are able to determine how an unsuccessful strategy correlates with a successful one, and how a student might change the former to achieve a more successful outcome. In many ways, this resembles the use of expert models while avoiding many of their pitfalls. In addition, unlike expert performances, these successful performances describe a wide range of strategies used by novice problem solvers, which allows the ANN to classify successive student problem solving attempts as these attempts move from novice-like to expert-like performances over time.

5.8.7 Training the ANN with “unitary” data
As described in Chapter 3, student performance data is presented to the ANN as a performance vector. As detailed there, the vector describes the order a student moved from one information item to another. My research with multiple IMMEX™ problems suggested modifications to the original data might be warranted, however. When training the Kohonen ANN previously described using fewer than 500 solved performances on IMMEX™ problems that had a large number of ways to solve each clone (like Desperately Seeking Solution), the resulting clusters were difficult to interpret. While the same student performances consistently clustered together, it was almost impossible for researchers or teachers to discern what made the student performances in each cluster similar.

Given that the information a student reviewed and the not order it was reviewed seemed most important, I reduced the detail of the information presented to the Kohonen ANN. I modified the data by substituting the value “1” for all non-zero item numbers in the data set. In effect, this replaces the order a student visited specific information items with the much less detailed fact that s)he merely visited the item sometime during his or her attempt at solving the problem. I then retrained the ANN. Both researchers and teachers agreed that the clusters produced by this data contained more similar strategies that were easier to discern and ultimately more useful for analysis of student performances than the previous results. In addition, the clusters produced by retraining the ANN in this way.
clearly identified various strategies the students used for solving the different clones of various IMMEX™ problems.

To ensure that the interpretation of each of the clusters was accurate, I created mock performances to duplicate the performance I felt each cluster most represented. Unlike the clusters produced by nets trained with fully weighted data, the ANN trained with this "unitary" data assigned these mock performances to the predicted cluster in every case. Nevertheless, a major drawback of training the ANN with this unitary data was the lack of consistency between training sessions. In this case, the same net trained with identical data on two separate occasions grouped slightly less than 70% of the performances it considered similar in the first round of training with those same performances in the second round of training.

In analyzing the performances which moved from one cluster to another between training sessions, it became apparent that the ANN required a larger amount of freedom in determining cluster boundaries than simple Kohonen unsupervised ANN allowed. In addition, it was also apparent that groups of performances, not individuals, generally moved from one cluster to another and that certain other performances seldom, if ever, moved between training sessions. Since these latter performances represented identifiable "anchor" clusters, they could be used as input to a supervised learning ANN to achieve better defined clusters.

When training supervised networks, the desired response of the network must be known for each training input. Ordinarily these would be determined a priori, the network would be trained and the trained ANN would merely classify the data into the appropriate cluster. However, because the unsupervised ANN had already identified performances which routinely clustered together, I used this knowledge to train supervised ANN that produced more defined boundaries around each cluster. The refinement: in cluster boundaries occurs because this supervised ANN can determine how far its current output is from its optimal output. Using the chain rule from the Calculus, the weight given each input, in this case the importance of each piece of IMMEX™ information a student did or did not choose, can be modified by the network according to its individual importance in the classification scheme for each cluster. The methodology employed here is endorsed by various authors. For example, Principie, et al. note that, "in many practical applications, data from each class tends to be dense, and there is a natural valley between classes. In such cases clustering can be a preprocessor for classification." (Principie, Euliano et al. 2000) Klahr and MacWhinney also suggest this arrangement "can increase the power of our models," and generate more satisfactory results. (Klahr and MacWhinney 1997) In fact, using the combination of unsupervised and supervised training, the reliability of clusters between training sessions (inter- "rater" reliability) improved to over 98%!

Although the techniques described above allow us to group individual student performances into general strategies, we must still develop a rubric to actually evaluate each strategy. Raizen cites the development of scoring protocols as the most difficult task in developing performance tasks. (Raizen 1990) Because of the problems cited
previously, it would be imprudent to merely compare students or other novices to their teacher or another expert. Not only do novices treat problems differently than experts (Messick 1989), but such comparisons are bound to limit the acceptable student approaches to problem solving. Nevertheless, the strategies used by the students are different from one another. Some of the strategies seem to routinely allow a student using the strategy to solve a problem, while other strategies almost never produce that result.

Consequently, for this research, I determine the effectiveness of student strategies by calculating the odds a particular strategy would produce the correct answer. While other metrics could be used, given the requirements of the national and state standards, along with those of Ms. Lindsay, student success rate seems an appropriate metric of strategy efficacy. Good strategies are those that tend to solve the problem. Furthermore, odds were used in place of a simple percentage measure to facilitate the comparison of strategies. Because the frequency of each type of clone a student received was a random event, comparing the percentage of time a strategy solved a particular clone was meaningless. For example, suppose a certain strategy was used by students to solve for the carbonate clones 5% of the time. Further suppose those students were successful 4% of the time and unsuccessful the remaining 1% of the time. If students used an alternative strategy to solve for the Carbonate clones 25% of the time and were successful 20% of the time, the comparison of 4% to 20% would be meaningless because the strategies were used with differing frequency. On the other hand, by computing a simple odds for each strategy (4%/1% in the first case, and 20%/5% in the second), one can see that, students are four times more likely to solve the problem than not using either strategy. As such, odds allow the comparison of solve rates between the various types of clones. For continuity an odds ratio was also used to compare overall solve rates for each strategy. This odds ratio is computed in the traditional manner (i.e. Probability of a correct answer / 1 – Probability of a correct answer). The fact that the natural logarithm of the odds a student solves a problem equals student ability (θ) in the Rasch Model, also made the odds ratio an obvious choice to facilitate other comparisons. (See Appendix 8).

It is important to stress that these rubrics should not be applied blindly in order to classify or evaluate individual students within a particular cluster. Obviously, individual students can use an ineffective strategy to solve IMMEX™ clones. In fact, while unlikely, the student could do so repeatedly even though the odds suggest such a strategy is ineffective. While rare, there are students in this data set that routinely solve IMMEX™ clones using ineffective strategies. For example, several students occasionally examined all available information before successfully solving a Hazmat clone. The odds of success for this kind of strategy are low for the population studied and, therefore make the strategy ineffective for a cohort; however, these several students routinely used such a strategy to solve the problem. As such, using strategy classifications to evaluate individual student performances against a criterion can be inaccurate and inappropriate if applied blindly. However, using strategy classification to help identify problematic learning and application is wholly appropriate. As the National Council on Testing and Public Policy observes, “the solution is not to avoid classifying people; such classifications are inevitable in modern society. It is to avoid classifying solely on the basis of one imperfect [measurement].” Our goal, as Messick observes, “is not to explain
any single isolated event, behavior, or item response, because these almost certainly reflect a confounding of multiple determinants. Rather, the intent is to account for consistency in behaviors or item responses, which frequently reflect distinguishable determinants.” (Messick 1989) To accomplish this requires investigating more than just the discrete, individual and fragile behavior of one student. Nevertheless, as will be shown, the use of ineffective strategies by a student usually does indicate that the student is struggling with the learning or application of content. Furthermore, without help the data suggests that a student will continue to struggle on similar IMMEX™ problems both in the near- and the long-term.

The effective use of generally “ineffective” strategies can also identify poorly designed clones. For example, if a student can routinely solve one or more clones without reviewing an adequate amount of information, the student may have prior knowledge about the problem of which the problem’s author was unaware. In a like manner, there may be “hints” in a problem that lead to the correct answer, but that only a few precocious students have identified.

“Effective” strategies must also be considered in the context of the performance which produced them. What is effective for one teacher’s students is not necessarily “effective” for another teacher’s students. By way of illustration, when I reviewed the performance of Ms. Lindsay’s students solving qualitative chemistry problems, I discovered they used virtually identical strategies to solve the problems that high school Advanced Placement (AP) and college freshman used to solve the same problems. The strategies used by the students were identical even though the students had different teachers, supposedly divergent academic abilities and encountered widely divergent curriculum. However, while one of the strategies was very effective for Ms. Lindsay’s students (using Litmus paper to test pH), it was less effective for the AP students and college undergraduates. These findings suggest that we can tease out “teacher effect” in various student performances. In fact, Ms. Lindsay indicated she had stressed the use of Litmus paper to measure pH in her classes, and one of the other teachers indicated she had placed greater stress on the reaction produced when adding an acid to a base. IMMEX™ seems to provide us a contextual sensitivity that previously cited test methods do not, and it allows stakeholders to determine whether students meet performance standards set for them even though the students may encounter differing academic programs.

A more important consideration when assessing individual student strategies is that, as expected, these strategies generally change in response to various clones. For example, a student solving for an unknown containing the hydroxide anion might only conduct a flame test and a litmus paper test before identifying the unknown. The same student might have to conduct five or six tests to identify an unknown containing the chloride anion. Consequently, the two strategies would fall into two distinctly different clusters. It is apparent in the data collected from Ms. Lindsay’s classes, that the particular strategies a student uses in one performance will not predict the particular strategy that student will use on a future performance because clone types are delivered randomly. Furthermore, as Messick recounts, numerous studies suggest that “different individuals perform the same task in different ways and that even the same individual might perform in a
different manner across items or on different occasions.” Messick refers to this as a “major conundrum” in educational and psychological measurement. (Messick 1989) The data from my research, however, suggest a way to begin addressing this conundrum.

The Hazmat data from Ms. Lindsay’s classes suggest a relationship between the amount of information a student looked at and the odds of solving a Hazmat clone using that information. As shown in Figure 5-3, if a student looks at very few pieces of information, her or his odds of correctly solving the problem are low; viewing too much information produces similar results.

**Figure 5-3.** This figure shows how the various types of strategies students used to solve Hazmat problems might be grouped. Strategies that relied on little information before trying to solve the problem had low odds of achieving a correct solution. These were termed limited strategies. Effective student strategies focused on the information most germane to the problem and ignored extraneous information. They had high odds of correctly identifying the unknown. Student strategies termed prolific explored large amounts of information (both relevant and irrelevant to the problem) and generally had low odds of achieving a correct solution to the problem.

If, however, a student identified particular information that was germane to the problem at hand, and limited her or his search to that information, the odds of solving the problem were high. This finding replicates the observations made by Hurst, Casillas, and Stevens
who report that the students who solve IMMEX\textsuperscript{TM} problems in the domain of medicine correctly access appropriate information and minimize search of unrelated concepts. (Hurst, Casillas et al. 1997) This suggests a more general classification of strategies might be appropriate.

Three distinct strategy “types” are apparent from the student performances in Ms. Lindsay’s class. Seven of the strategies identified by the ANN reveal that the students view very few pieces of information before attempting to solve the problem. In fact, five of these strategies do not investigate enough information to solve any of the Hazmat clones. These strategies can be referred to as “limited” strategies. Conversely, eight strategies indicate that students are viewing a great deal of information before attempting to solve the problem. Although these strategies indicate that the students often viewed the information necessary to solve the problem, they also reveal that students routinely accessed information that was irrelevant and uninformative to the solution of every Hazmat clone. These strategies can be termed “prolific” strategies. Students using limited or prolific strategy types generally have less than even odds of solving Hazmat clones. On the other hand, students using the remaining strategies are all more likely than not to correctly solve the problem posed to them. Because these strategies focus almost exclusively on examining the information necessary to solve the problem at hand, I classify them as “efficient” strategy types.

In addition to the odds of solving a problem, other metrics suggest a similar grouping of strategies under each strategy type. When evaluating individual student performances, it becomes evident that the student either solved the problem or did not solve the problem, and that they either had enough information to solve the problem or did not have the required information to solve the problem. These possibilities allow for four possible classifications. If a student had enough information to solve the problem and did so I considered them to have demonstrated “understanding.” If a student had enough information, but did not solve the problem, I termed the performance “algorithmic”. I use this term to imply that while the student may have known what information to look at, they apparently did not know how to interpret the information they observed. Conversely, if a student did not have enough information to solve the problem, but, in fact, solved it, that performance is termed “guessing”. Finally, if a student did not have the information required to solve the clone they were attempting and did not solve it, I classified the strategy as “unable to progress” toward a solution. In virtually every case, the “limited” strategy type encompassed virtually all the student performances classified as “guessing” or “unable to progress.” Similarly, the “prolific” strategy type enveloped the student performances classified as “algorithmic”. Finally, and not unexpectedly, nearly all performances demonstrating “understanding” fell into the “efficient” strategy type. I will explore the validity and use of the “understanding” construct in greater detail in the following chapter.

5.9 Predicting student ability

Because these strategy types are less constrained than any one of the strategies they contain, it was expected that once a student had demonstrated the ability to use an
"efficient" strategy, s)he would generally continue to use "efficient" strategies on subsequent problems.

To investigate this hypothesis, I compared the strategy type of the student’s first performance to the strategy type the student used most often in solving Hazmat problems. The results are shown in Figure 5-4.

<table>
<thead>
<tr>
<th>Hazmat Problem</th>
<th>&quot;Limited&quot; Modal Strategy</th>
<th>&quot;Efficient&quot; Modal Strategy</th>
<th>&quot;Prolific&quot; Modal Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Limited&quot;</td>
<td>30</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>First Strategy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Efficient&quot;</td>
<td>5</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>First Strategy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Prolific&quot;</td>
<td>5</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>First Strategy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

($\chi^2 = 70.5; df = 4; p < .001$)

Figure 5-4. This figure shows the relationship between the type of strategy a student used to solve their first Hazmat clone and the strategy the same student used most often (the mode) to solve subsequent Hazmat clones. Overwhelmingly, the strategy type used by a student in his or her initial performance predicts the strategy type they will continue to use for future Hazmat problems. The chi-square statistic suggests that this is not a random occurrence.

Overwhelmingly students that used efficient strategies on their first performance continued to use these same type of strategies on succeeding attempts to solve Hazmat problems. Unexpectedly however, students who used less optimal types of strategies continued to repeatedly use these same types of strategies in future problem solving. Apparently, even though the students were unlikely to solve Hazmat clones using a "limited" or "prolific" strategy, without teacher intervention they continued to use that same type of ineffective strategy on future problems. This occurred even though the students had immediate feedback about whether they had solved each clone, and knew they would be evaluated on how they had solved the problem.

This finding is important not only because it demonstrates the concurrent criterion validity of the IMMEX™ Hazmat problem. Indeed, subsequent research has identified this same relationship in other IMMEX™ problems and other classrooms as well. Beyond that, however, this finding suggests that teachers can not only use strategy types to identify students who are likely to have trouble applying content knowledge to specific performance tasks, but also as an indicator of the effectiveness of interventions which could improve student learning. Pask, as cited in Messick, held that such systematic errors represent pathologies in learning and performance. Messick adds that "one of the most useful approaches to the understanding of human functioning is to study the ways in which it breaks down or goes awry." (Messick 1989) My research suggests that the IMMEX™ tool can aid educators and researchers in identifying such dysfunctional pathologies. Moreover, by studying the performance of those students who transition from a "limited" strategy or from a "prolific" strategy to generally using an "efficient" strategy, the performances of the students themselves might suggest specific, contextually sensitive pedagogical interventions an individual teacher might make to improve the
learning of his or her students. Similarly, the transitions of a student's performance from "efficient" to one of the less efficient strategy types may suggest other interventions to improve instructional quality for a particular student or group of students.

As shown in Figure 5-5, this same trend also seems to be developing in the student performances in *Desperately Seeking Solution*. However, for this problem, more of the student's performances are classified as "Limited" strategies. As mentioned in section 5.5, I believe this arises due to the fact that the students only conducted about half the performances using *Desperately Seeking Solution* as they did with *Hazmat*. This large change in performances can have a dramatic effect on the odds that students will solve a problem as overall solution rates tend to be lower when students initially begin to solve problems. Because the odds of solving a problem are integral to the determination of "strategy types" (See Figure 5-3), when that axis is restricted, the categories tend to show less distinct boundaries. Furthermore, *Desperately Seeking Solution* has almost four times as many pieces of information as *Hazmat*. I suspect many students may not have taken the time necessary to find the information required to correctly identify the unknown.

<table>
<thead>
<tr>
<th>DSS Problem</th>
<th>&quot;Limited&quot; Modal Strategy</th>
<th>&quot;Efficient&quot; Modal Strategy</th>
<th>&quot;Prolific&quot; Modal Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Limited&quot; First Strategy</td>
<td>56</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>&quot;Efficient&quot; First Strategy</td>
<td>10</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>&quot;Prolific&quot; First Strategy</td>
<td>6</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

\[\chi^2 = 64.4; \text{df} = 4; p < .001\]

**Figure 5-5.** This figure shows the relationship between the type of strategy a student uses to solve their first *Desperately Seeking Solution* clone and the strategy the same student uses most often (the mode) to solve subsequent *Desperately Seeking Solution* clones. Overwhelmingly, the strategy type used by a student in his or her initial performance predicts the strategy type(s) he will continue to use for future *Desperately Seeking Solution* problems. The chi-square statistic suggests that this is not a random occurrence. The large number of students using limited-strategies may be due to limited student experience with this IMMEX™ problem or because the students performed a limited number of clones making strategy classification more difficult.

As discussed in Chapter 2, short term (concurrent) prediction is only one facet of determining how accurate our inferences of student performance are when measured against a set of criteria. We must also demonstrate the ability of a measurement instrument to produce results that accurately demonstrate how a student will perform at some future time (predictive ability). In fact, unlike percent, PI or IRT thetas, student performance on the *Hazmat* clone does suggest similar performance on the *Desperately Seeking Solution* clones that students attempted to solve near the end of the semester. As can be seen in Figure 5-6 below, without pedagogical intervention the general strategy type a student used to solve *Hazmat* problems is reflected in the strategy types students will use to solve clones in *Desperately Seeking Solution*. 

118
<table>
<thead>
<tr>
<th>Used “limited” strategy to solve DSS</th>
<th>Used “efficient” strategy to solve DSS</th>
<th>Use “prolific” strategy to solve DSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used “limited” strategy to solve Hazmat</td>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>Used “efficient” strategy to solve Hazmat</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Used “prolific” strategy to solve Hazmat</td>
<td>25</td>
<td>8</td>
</tr>
</tbody>
</table>

($\chi^2 = 11.86; df = 4; p = .018$)

**Figure 5-6.** This figure shows the relationship between the type of strategy a student generally uses to solve Hazmat clones and the strategy the same student used most often to solve Desperately Seeking Solution clones. The strategy type used by a student in Hazmat often predicts the strategy (s)he will use to solve Desperately Seeking Solution problems. The chi-square statistic suggests that this is not a random occurrence. Here again, the limited number of Desperately Seeking Solution problems performed by the students have probably over-inflated the number of strategies classified as “limited” in Desperately Seeking Solution.

While Figure 5-6 does indicate a large number of students adopt limited strategies to solve *Desperately Seeking Solution* after using prolific strategies on *Hazmat*, this seems to result from the classification of *Desperately Seeking Solution* strategies as explained previously. Nevertheless, the trend of students to adopt and keep a particular strategy is also clear in this data. The modal strategies adopted by the students correlate with other measures to varying degrees; however, those who use limited strategies tend to have the lowest grades (GPA, first and second semester chemistry, and chapter test). They are also more likely to have lower SAT-9 scores (on all tests), and to evaluate their own ability among the lowest in the class. The rest of Ms. Lindsay students are approximately randomly scattered between the “prolific” or “efficient” strategy categories.

There are no gender or SES correlations with a student’s most often used *Hazmat* strategy type. While there is no significant difference in the ethnicity of students who use “efficient” strategies, non-white students tended to disproportionately favor “limited” strategies, while white students used “prolific” strategies more often than would be expected in a random distribution.

**5.10 Conclusions**

Assessing what someone knows or can actually do requires standards. Furthermore, these standards must be understandable, acceptable, realistic, and public in order affect changes in knowledge or behavior. While two schools of thought exist (task analysis and the Tyler approach), the de facto national standards developed by the NRC, the California science standards, and the standards used by Ms. Lindsay all have a definite behavioral component. These current standards favor the Tyler approach’s evaluation of both content and behavior.
Unfortunately, the ubiquitous evaluation tools presently in use are ill-suited to evaluate student learning using these criteria. First, the national and state standards are both entirely voluntary so no assessment tool is required to adhere to these standards. As a consequence, the criteria against which student learning is measured are somewhat arbitrary and can produce very misleading results if they are interpreted in light of the present standards. Perhaps the best example of this is the discontinuity between NEAP results published by the Department of Education and the results of various state tests for a virtually identical population of students. Without a clear standard against which to gauge such results, the inferences made about student learning from the data these tests provide are meaningless. One can make a similar argument for the meaning of student grades. While there appears to be general agreement on what an “A” or a “C” means, there is no guarantee that a grade from one teacher is the same as an identical grade from another teacher. Without a viable criterion, no one knows what the results mean.

Second, without understandable, acceptable, realistic, and public guidance, what gets tested and taught is determined by individual teachers. Many educational stakeholders are excluded. As previously discussed, when high-stakes tests are introduced into such an environment, actual curriculum generally devolve into “teaching to the test.” Unfortunately, in the US system of education, this has often meant a sub-standard education for many students, especially those in urban schools. When individual standards are also unspoken or hidden, the research suggests that the result is often ineffective learning. Learners need to know what is expected and how to evaluate their progress toward those expectations in order to effectively learn.

While the national and state standards in place during this research were not binding on the students studied, Ms. Lindsay was clear in her standard of what she wanted her students to do, and made her criteria clear to her students when they began working IMMEX™ problems. Briefly, she wanted them to apply their qualitative chemistry knowledge in authentic situations in order to correctly identify an unknown compound. Furthermore, she wanted them to write their thinking down on worksheets in order to demonstrate the logic they used to arrive at a conclusion. The students knew that they would have similar practice and assessment problems, and that both would be assessed by the teacher. In both cases, the logic a student used to eliminate or confirm an unknown would be evaluated. As such, Ms. Lindsay was interested in both the content of a student’s work as well as how each actually applied that content to the problem at hand.

Although Ms. Lindsay felt that standard pencil and paper tests allowed her the opportunity to ascertain a student’s content knowledge, she did not believe they afforded her the ability to evaluate how well a student could apply that knowledge. While the NRC and California standards differ on the degree to which students will be able to apply what they know, both suggest this is an important aspect of what a high school chemistry student should be able to do.

Technological advances offer us the opportunity to assess students in new ways. IMMEX™ technology, in particular, allows the stakeholders in the educational
enterprise, especially students and teachers, to observe not only what a student answered, but how a student arrived at that answer. While evaluating what a student answered can suggest when students are struggling with material, trends in learning, or even when assessment should occur, the inferences we can make from such data is limited. The research from this study suggests that if we don’t exploit the data generated by tools like IMMEX™, neither the insights suggested by our inferences nor the validity of those inferences will dramatically improve. Current measures such as student ability as measured by Item Response Theory, or measures of percentage correct did not predict near or long term student success or failure. Arguably this is because these metrics are not robust enough. They measure if, not why, a student was able to solve a problem. While such indicators could suggest if a student possesses basic knowledge, they measure the ability to actually apply that knowledge in a very tangential way.

The amount of data IMMEX™ generates about a student performance, however, can be overwhelming to a human assessor. For our purposes, there seem to be almost an infinite number of ways a student may solve the problems presented to them. Again, technology offers us the tools to make sense out of these performances. This chapter demonstrated that supervised and unsupervised artificial neural networks could discern patterns in actual student performances to aid our understanding of how students actually applied content knowledge to solve a problem. This methodology is fundamentally different from current methods in educational assessment in that students are not forced into a model of learning that is defined a priori. Rather, the actual student performances form the basis of our inferences about student problem solving.

My research did confirm that students of different ability levels, classes, and schools use the same strategies to solve problems; however, I also found that these strategies are not always as effective in one class as they are in another. Undoubtedly the effectiveness of a strategy involves teacher effect. In at least one case, IMMEX™ allowed us to both discern and begin to explain that effect. This alone suggests that the inferences we can begin to make with these tools can significantly advance our inferences about student learning.

In this work, the effectiveness of a student performance is measured by the odds a strategy would work, in a given context, to solve the problem at hand since successful problem solving is implicit in the current standards. This method does not compare the students in one population to those in another similar or dissimilar population in order to determine effectiveness. The research suggests that student strategies, in both Hazmat and Desperately Seeking Solution, fall into one of three broad category types. Strategies that investigate small amounts of information before attempting a problem solution generally have low odds of success. These were termed limited. Similarly, strategies that involve the student in looking at a plethora of information before trying to solve the problem also have low success rates. Finally, in strategies where the student limits his or her search to a few important pieces of information, the odds they will solve the problem are high. Further research has suggested the same pattern for other IMMEX™ problems as well. Using such a categorization suggests we can begin to unravel Messick’s “educational conundrum” because we can account for students using various strategies at
different times. More importantly, the research suggests that, without teacher intervention or more directed feedback, students will continue to use the same efficient or inefficient strategy they use on their first clone over and over again on subsequent clones. Unexpectedly, this occurs even if the student knows s/he is often unsuccessful at solving the problem using such a strategy. The research here also found that these trends are not only significant between versions of the same problem, but between similar problems as well.

The significant correlation between the strategy type a student uses to solve her or his first and subsequent problems argues strongly for both concurrent and predictive criterion validity of the inferences from IMMEX™ data. In hindsight however, the most compelling result of these findings suggests the tool could be an effective way for teachers to identify students who are having difficulty learning or applying content knowledge in a practical manner. Unlike present assessments, IMMEX™ offers an accurate way to assess both the content and the behavioral aspects of student learning.

Although comparing a student to some criteria may allow us to infer that the student will perform similarly in the near or long term future, the criteria is not intended to explain why the student performed in a certain manner. In other words, I have yet to demonstrate that because a student consistently used “efficient” strategies they understand qualitative chemistry. While the reader might recall that understanding, as described in this chapter, was apparently related to the use of “effective” strategies, that correlation was not explored in the present chapter. Despite the fact that the two may be related, demonstrating that a particular student possesses a particular psychological trait is the subject of construct validity. It is to this type of inference validity that we now turn.
Chapter 6. Inferring why students can do what they do

The need to accurately assess a student’s ability to learn and apply the content of a specific domain to authentic problems associated with that domain has been a recurrent theme in previous chapters. I have argued accurate inferences about student achievement require that we specify and assess a specific domain of knowledge and that we articulate appropriate criteria against which to measure student achievement. I have also cited research that suggests that real-world or authentic assessment tasks not only allow more accurate assessment, but can have significant positive effects on our educational system.

In each of these cases, however, we were measuring tangible indicators of student learning. When designing assessments to match the curricular and instructional content of a course, we were able to measure a question or concept from the course and compare that with questions or concepts measured on an assessment. Experts were able to review specific questions and we could use these reviews to decide the appropriateness of questions. Subsequently, the national, state and classroom standards allowed us to gauge how well students could demonstrate a level of content mastery and their ability to apply that knowledge to a specific task. Both Ms. Lindsay and I were able to examine a student’s performance to see if it satisfied the criteria of a particular standard. Using some tools afforded by modern technology, I demonstrated how we can predict near- and long-term student performance on similar problems. There is, as a result, strong, statistically significant evidence to conclude that IMMEX™ technology allows us to measure both how well a student knows the content in a specified domain, and how well a student can apply that knowledge to authentic problems from that domain. However, merely inferring that a student can appropriately apply the content from a finite knowledge domain to a problem within that domain generally forms only the most basic of the inferences we desire to make about students. More often, we want to know what the student’s performance means and discern why the student is able or unable to do what they do. In other words, we want an idea of what trait of the student explains her or his ability or lack of ability in a certain domain.

Unlike the breadth and depth of content or the ability to replicate a performance, traits are not tangible and so cannot be directly observed or measured. In educational and psychological testing, these intangible traits are often referred to as constructs (APA 1985) because we often build them from other factors. We confirm the presence of the construct by confirming the presence of the factors. In order to measure these intangible constructs then, we must rely on the “product” a student produces or the actions a student takes to infer that the factors, and therefore the construct is, in fact, present. (Bloom, Madaus et al. 1981) Because our ability to infer the presence of such intangibles requires accurate measurement of more tangible artifacts and because it relies on both content and criterion validity, construct validity is often referred to as the “overarching” (Cronbach 1971), or “super-ordinate” validity type. (Anastasi 1990) In fact, Messick suggests that construct validity is ultimately the only measure of the accuracy of our assessments. (Messick 1989) While the content assessed must be relevant and representative, and it must be accurately measured against specified criteria, these forms of validity are
necessary, but are not sufficient to assure one can accurately evaluate the trait or traits that *explain* the observed behavior or product.

The trait of intelligence is one of the most classic constructs in the field of educational testing. As Chapter 2 discussed, our attempts to measure intelligence not only formed the basis of the present system of educational measurement in the United States, but continues to occupy research in the field to this day. (Duncan, Seitz et al. 2000) More recently, the trait of understanding has become increasingly important to policy-makers and other stakeholders in the U.S. educational system. We want to infer not only how well students know and use the content in a particular domain, but also how well a student can “use their mind” (*Goals 2000: Educate America Act*). We want to infer student understanding.

### 6.1 Defining the construct of “understanding”

Since understanding can be defined in different ways, it is important we define such a construct before trying to measure it. Webster’s dictionary defines understanding as the ability to perceive *meaning* and suggests that one who understands can grasp the *significance* of that which is understood. Brady refines the definition by suggesting that one who understands is familiar with the *relationships* between concepts. Furthermore, she suggests that when educating for understanding, we must be concerned with the construction, organization and elaboration of related ideas or concepts. (Brady 1989) Perkins defines understanding even more specifically. He suggests understanding can be identified by three characteristics. He proposes that those who understand a topic can: 1) satisfactorily explain that topic or concept to another; 2) support their explanations by a “web of facts” so that their explanations are related to other knowledge that s)he can access; and 3) revise and extended their explanations as new knowledge is uncovered or learned. While Perkins suggests that students must “learn the rules of a domain”, he also observes that problem solving is needed to demonstrate understanding. (Perkins, Crismond et al. 1995)

Although other measures must be used to impute the existence of constructs, this does not mean that *any* observable can be measured and used to infer the existence of some construct. Unfortunately, this requirement has not always been well understood. The history of intelligence testing produced some disturbing examples based on prejudice and bias. To ensure the credible and accurate measure of constructs, Cronbach and others stress the need to first support the relationship between a construct and the observables one will use to measure that construct by demonstrating that a logical theory connects the two. This theory should not only posit a relationship between the construct and what is observed, but should explain why that relationship exists. (Cronbach 1971; Messick 1989; Shepard 1993)

Using the definitions of understanding suggested above, a student who understands the domain of qualitative chemistry should be able to explain how they have used their content knowledge to identify an unknown chemical compound. They should be able to identify, interpret and use appropriate and necessary facts to reach a conclusion and be
able to eliminate or discount other reasonable conclusions. As such, the amount of information a student uses to reach a conclusion is not as important as is the fact that the information they use justifies their conclusions. Finally, they should be able to extend or revise their hypothesis or conclusion as new information makes such changes necessary. In theory, those who are able to demonstrate such behaviors can be said to understand because they synthesize interpreted facts to reach a logical and valid conclusion. Furthermore, since the interpretation of the facts, rather than a memorized algorithm, is required to form an answer, the student who understands an answer will adapt his or her hypothesis to new or unexpected information.

Perkins’ definition makes intuitive sense. One who understands must be able to do more than merely recite knowledge or apply it in a mechanical way identical to the way it was learned. The former suggests mere memorization, the latter training. In neither case could the argument be made that a student understands. Bloom’s taxonomy suggests that to go beyond knowledge and demonstrate understanding, students should be able to utilize what is known in new situations or in “a modified form from that in which it was originally encountered.” (Bloom 1956) Unfortunately, he also laments that much of what was occurring in the classrooms of his day was the rote memorization and assessment of knowledge. Much of the research I have previously cited suggests that these activities typify current classrooms and assessments as well. Nevertheless, research also suggests that new pedagogical styles and the proper application of technology can affect an educational system that teaches for understanding.

6.2 Measuring “understanding”

I measured student understanding in two ways in this research. First, in IMMEX™ problems understanding can be measured by determining how a student solved a problem. Secondly, by using the notes student were directed to make while working the Hazmat and Desperately Seeking Solution problems, Ms. Lindsay can measure how the student justified the way they solved the problem. As was developed in Chapter 3, the IMMEX™ technology presents a problem to be solved and the information necessary to solve the problem. A student may access any or all information in any order to solve the presented problem. I assume that students will continue to access information until they believe they can solve the problem. It is conceivable that this might not be the case in all IMMEX™ performances (e.g. when students were under time constraints to finish solving a problem before they were ready). However, since the IMMEX™ performances I use in this part of the research were generated by students being assessed in the classroom, there were no such limitations on their performances. The students also had their grade as a motivation to demonstrate they could correctly solve the problem they were presented. As such, I assert that the information each student reviewed forms an explanation of how s)he solved the problem they worked on.

In both the qualitative chemistry problems used for this research, there is enough information to solve the problem in multiple ways, and there are informational items that move a problem-solver toward a solution. Because information must be interpreted in order to reach a conclusion, merely memorizing a series of steps necessary to solve one
problem will not necessarily allow one to solve the problem presented. The fact that various clones requiring different information to solve are presented to a problem-solver in random order further reduces the chances that one can merely memorize a series of steps and solve the clone without understanding or making meaning of the information presented.

Arguably, a student could memorize a series of steps, correctly interpret the specific information generated by taking such a path, and ultimately solve the problem. However, this seems a perfect illustration of the “fragile” understanding Perkins suggests. (Perkins, Crismond et al. 1995) For example, in observing a student solve Hazmat problems, I watched as she conducted a flame test followed by a test with litmus paper. In this case, the student not only correctly identified the unknown, but, having correctly interpreted the tests, she knew that she could stop gathering information because she had eliminated all other possibilities in the data set she had. Subsequently, when she encountered an unknown that did not contain the hydroxide ion, we would logically expect her to not only know that the unknown could not be hydroxide, as the Litmus test would have produced the “incorrect” results, but also that she must keep gathering information before she could conclusively identify the unknown. This would have been true even if she did not correctly identify the second unknown. She, in fact, “understood” this second compound was not a compound containing the hydroxide ion. Logically, the more diverse the types of problems a student can solve and the more often s/he can solve problems, the greater the degree of understanding the student demonstrates. Nevertheless, while more information suggests the degree of understanding, even the student’s first attempt indicated some understanding.

By its very nature, IMMEX™ is a problem solving tool that requires students to identify, interpret, and use facts to logically solve problems in a particular domain. While there often is no one way to reach the solution to a problem, it is an easy task to ensure that students use enough information to both eliminate all other possibilities and confirm the correct answer.

As mentioned in Chapter 5, the students solving IMMEX™ problems were classified into four construct categories. Only the one termed “understanding” was investigated in this study. As I have suggested, a student collecting enough information to eliminate all but one or confirm only one answer and who answered the problem correctly was considered to have demonstrated understanding. Consequently, a very simple indicator of understanding is that the student has enough information to solve the problem and the student actually uses that information to solve the problem. Using such a metric, the student is able to explain their understanding by solving the problem correctly, they choose the information necessary to confirm the answer to the problem and, when necessary, keep exploring information until they can conclusively identify an answer. In this way, they incorporate new information. Finally, although the students had seen each ion of the clone, the students performing Hazmat were tested on actual clones they had never experienced so the assessment problem was a new situation.
This classification scheme fits well with Perkins' and Bloom's definition of understanding. For the purposes of this research a student is said to understand if s)he can:

- "Explain" the topic to another person (the researcher) by correctly solving the problem with information gathered from the problem space.
- Use facts to support a solution of the problem presented by ordering sufficient information to eliminate all but one or confirm only one solution.
- Revise their explanation of the topic to incorporate new knowledge by adapting their strategy to solve for dissimilar unknowns, and knowing when additional tests are needed to confirm an unexpected unknown.
- Use their explanation in new, unlabeled situations. In this study the specific unknowns used for the qualitative chemistry assessment had not been practiced by the students. Although the students had practiced with each ion presented, they had not practiced with the combination of ions presented during assessment.

While there is a slight chance that this scheme will misclassify a student as understanding when they, in fact, do not, the chances are smaller than might be expected. First, as indicated in Chapter 5, most students who order more tests than necessary to solve the problem seldom correctly identify the unknown compound. Therefore, students who order a lot of extraneous information are unlikely to solve the problem and are unlikely to be classified as having understanding. Moreover, this method of classifying students does not penalize the perfectionist students who want to make absolutely sure they have correctly identified the unknown by examining all available information. Secondly, as the clones change from one attempt to another, the amount of information required to reach a solution changes. Therefore, students who don't adapt their strategies to each new situation are unlikely to have enough information to solve new problems and, therefore, are unlikely to be classified as having understanding.

Nevertheless, some students classified as understanding will not, in fact, understand (i.e. false positives). Likewise, there is a chance that students who do understand will be classified as not understanding (i.e. false negatives). In addition to the misclassifications discussed above, the former could also occur when a student checks the information necessary to solve the problem, doesn't understand, and then randomly picks a correct solution. The latter could occur when a student chooses the appropriate information and correctly interprets it, but inadvertently chooses the incorrect answer. While any classification of human behavior is, by its nature, imperfect, as I will demonstrate the correlation between this schema and another more traditional measurement of student understanding is large enough to suggest misclassification is not a serious threat here.

Messick argues that to substantiate a relationship between the construct and the metric used to measure it, one must "appraise the degree to which empirical relationships with other measures, or the lack thereof, are consistent with that meaning." Furthermore, he suggests that these measures should not be tied to any particular mode of measurement.
(Messick 1989) Obviously, if IMMEX™ and the teacher are, in fact, measuring the same construct, the two should correlate significantly.

For this study, Ms. Lindsay created her own five point rubric to measure student understanding. The rubric was structured as follows:

- Five points: A performance that explicitly shows the logic behind the student’s correct answer. All other possible answers are eliminated.
- Four points: A performance that shows the logic behind a student’s answer, but the logic required a student to guess at least twice before arriving at the correct answer.
- Three points: A performance that gave the information necessary to solve the problem, but the student appeared unable to interpret the information and arrive at a conclusion.
- Two points: A performance in which the student began a series of tests (e.g. identified part of the unknown) or made one or two initial observations.
- One point: A performance that was apparently random or showed little logical direction.

Ms. Lindsay used this rubric to evaluate her students’ Hazmat performance worksheets (Appendix 3) at the end of the qualitative chemistry unit. As was the case above, she wanted to assess how well each student understood the concepts in qualitative chemistry. She focused on how well the student explained the logic they used to eliminate other possibilities and confirm an answer, if the facts allowed them to draw such a conclusion, and if the students assessed the necessary information to solve that particular Hazmat clone.

Because the IMMEX™ based metric and the teacher rubric were arguably measuring the same construct of understanding, one would expect them to be highly correlated. In fact as Figure 6-1 suggests the two metrics seem to be measuring the same things:

---

8 For grade weighting purposes, Ms. Lindsay used a 10 – 50 point scale but assigned grades in increments of 10 points. For ease of use, I converted this to a 1-5 point scale.
<table>
<thead>
<tr>
<th>IMMEX™ Classification</th>
<th>Rubric Score “1”</th>
<th>Rubric Score “2”</th>
<th>Rubric Score “3”</th>
<th>Rubric Score “4”</th>
<th>Rubric Score “5”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>37</td>
</tr>
<tr>
<td>Did not fully Understand</td>
<td>1</td>
<td>23</td>
<td>16</td>
<td>18</td>
<td>8</td>
</tr>
</tbody>
</table>

$\chi^2 = 52.38; \text{df} = 4; p < .001$

**Figure 6-1.** This figure is a cross-tabulation of the score a student received on the notes (s)he made to support his or her answer on the Hazmat assessment problem versus how the student’s performance was classified using their IMMEX™ performance alone. If a student accessed enough information to solve the problem and actually solved it, the student was classified as having demonstrated “understanding.” Student performances evaluated as a “5” by Ms. Lindsay demonstrated a similar level of “understanding.” The number of students in each category is shown. The chi-square statistic suggests that the distribution did not occur by random chance.

Figure 6-1 suggests there is a significant correlation between students classified as understanding based on their IMMEX™ performance and those so classified by Ms. Lindsay’s rubric. In fact, the actual data suggests this correlation may be even stronger than indicated in the figure. In half of the eight performances that received a score of “5” on Ms. Lindsay’s rubric and were classified as “Did not fully Understand” based on the student performance recorded in IMMEX™, the student never actually attempted the problem they indicated on their worksheet. Rather s)he merely made up a non-existent performance on paper so s)he would not have to actually work the clone that was presented by the computer. Usually this “invented” performance replicated a successful practice performance from earlier in the semester. The large number of assessment performances made these faked performances almost impossible for Ms. Lindsay to detect without matching the paper and computer records. In this case, the structure of the study prevented her from doing so. Furthermore, the 12 performances classified as demonstrating understanding but given scores of less than five by Ms. Lindsay usually included more tests in IMMEX™ than their note pages actually indicated or the student did not indicate the logic used to arrive at an answer on their note pages. For example, the students did not show all the information they looked at or did not justify their conclusions with ionic equations.

In addition to confirming that the two metrics are measuring the same construct, it is also important to ensure neither shows an unexpected relationship with an unrelated construct. Messick and Cronbach both suggest that the correlation between metrics that measure the same construct should be higher than the correlation between either metric and other measures of different constructs. In Cronbach’s words, “what the construct measure doesn’t correlate well with is just as important as what it does correlate to.” (Cronbach 1971; Messick 1989) Figure 6-2 suggests that the two measures of understanding have a much greater correlation with one another than either metric has with a content evaluation such as the Unit 12 written qualitative chemistry test, the student’s self-evaluation, and the student’s self-reported frequency of guessing at a solution.
<table>
<thead>
<tr>
<th></th>
<th>IMMEX™ Understanding</th>
<th>Teacher Rubric</th>
<th>Unit 12 Test</th>
<th>Student Self-evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Rubric</td>
<td>.638</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit 12 Test</td>
<td>.295</td>
<td>.397</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student Self-Evaluation</td>
<td>.303</td>
<td>.375</td>
<td>.273</td>
<td></td>
</tr>
<tr>
<td>Student self-reported</td>
<td>-.302</td>
<td>Not Significant</td>
<td>-.256</td>
<td>-.374</td>
</tr>
<tr>
<td>guessing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6-2.** The first diagonal of Cronbach’s correlation triangle suggests that, in fact, the two measures of “understanding” demonstrate the highest degree of correlation. Furthermore, neither measure correlates as well with metrics measuring content (Unit 12 test) and guessing (Student self-reported guessing). All correlation coefficients are significant at $p < .01$.

This suggests not only that the metric called IMMEX™ understanding and the metric called Teacher rubric are measuring the same construct, but also that neither of these metrics is actually measuring another specific construct like content knowledge or guessing. In addition to the appropriate correlation coefficients in Figure 6-2, the coefficients recorded in Figure 6-3 suggest a similar relationship between both the IMMEX™ understanding and Teacher rubric metrics and high stakes content tests. Here again, correlation coefficients less than .638 suggest that neither metric of understanding is actually measuring only the content knowledge reputedly tested by grades or the state test. The lack of a correlation between the measures of understanding and SAT-9 reading suggests that, unlike grade measures (see Figure 6-4), both understanding metrics are unbiased by a student’s English language literacy. These results suggest that these other scores appear to be measuring something other than what is being measured by the construct we have termed understanding. Similarly, the construct understanding, as measured both by Ms. Lindsay and the student’s IMMEX™ performance, is not correlated with a student’s gender, ethnicity, or socioeconomic status.
<table>
<thead>
<tr>
<th></th>
<th>IMMEX™ Understanding</th>
<th>Teacher Rubric</th>
<th>Unit 12 Test</th>
<th>Student Self-evaluation</th>
<th>Student self-reported guessing</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>.283</td>
<td>.368</td>
<td>.545</td>
<td>.248</td>
<td>-.350</td>
</tr>
<tr>
<td>1st Semester chemistry</td>
<td>.205 *</td>
<td>.358</td>
<td>.637</td>
<td>.281</td>
<td>-.249</td>
</tr>
<tr>
<td>2nd Semester chemistry</td>
<td>.246</td>
<td>.437</td>
<td>.580</td>
<td>.216 *</td>
<td>-.245</td>
</tr>
<tr>
<td>SAT-9 reading</td>
<td>Not Significant</td>
<td>Not Significant</td>
<td>.456</td>
<td>Not Significant</td>
<td>Not Significant</td>
</tr>
<tr>
<td>SAT-9 math</td>
<td>Not Significant</td>
<td>.236</td>
<td>.352</td>
<td>.325</td>
<td>Not Significant</td>
</tr>
<tr>
<td>SAT-9 science</td>
<td>Not Significant</td>
<td>Not Significant</td>
<td>.437</td>
<td>.293</td>
<td>Not Significant</td>
</tr>
</tbody>
</table>

**Figure 6-3.** This figure shows the correlation coefficients between the two measures of understanding (IMMEX™ and Teacher rubric) and student evaluation, student reported guessing, and other pencil and paper evaluations. As expected, the pencil and paper evaluations demonstrate the greatest correlations with measures of content “knowledge”. In addition, the measures of understanding show a much lower correlation with content measures. All coefficients are significant at p < .01 unless noted. The coefficients with an asterisk are significant at .01 < p < .05.

<table>
<thead>
<tr>
<th></th>
<th>GPA</th>
<th>1st Semester chemistry</th>
<th>2nd Semester chemistry</th>
<th>SAT-9 reading</th>
<th>SAT-9 math</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Semester chemistry</td>
<td>.673</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Semester chemistry</td>
<td>.757</td>
<td>.747</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-9 reading</td>
<td>.465</td>
<td>.414</td>
<td>.311</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-9 math</td>
<td>.288</td>
<td>.356</td>
<td>.237</td>
<td>.336</td>
<td></td>
</tr>
<tr>
<td>SAT-9 science</td>
<td>.291</td>
<td>.363</td>
<td>Not Significant</td>
<td>.642</td>
<td>.540</td>
</tr>
</tbody>
</table>

**Figure 6-4.** This figure shows the lower diagonal of Cronbach’s correlation triangle. As expected the inter-correlation coefficients between grades and between parts of the Stanford Achievement Test, 9th edition are generally large indicating they are measuring similar constructs. Furthermore, the correlation between grades and SAT-9 reading suggests that English language literacy may play an important role in grade evaluations. All correlation coefficients are significant at p < .01.

Unlike standard pencil and paper tests, IMMEX™ technology gives us the ability to observe students as they explain the way they would solve the problem. We are able to evaluate strategies and performances based on how likely students using such a strategy are to solve the problem they are working on. This is critical to an accurate assessment of student performance, for as Maker notes, often answers teachers think are wrong or “inappropriate because they fall outside the traditional way of viewing the solution may be a richer source of information about students than the product we expect.” (Maker 1994)
In fact, IMMEX™ allowed me to observe just such an occurrence during the course of this study. While evaluating a group of AP student solution strategies, I noticed a strategy cluster that suggested the strategy would probably be very inefficient because the students relied on so many tests. In fact, the odds of solving the problem using such a strategy were quite high (more than 3 to 1). Upon further exploration, I discovered the students had used a novel strategy by combining their understanding of physical and chemical properties. The student’s teacher, who is one of Ms. Lindsay’s peers, was quite surprised when I reported my findings to her, as she never taught such a strategy in her class. Apparently, the students had synthesized the strategy from content they had been taught during the course. More importantly, IMMEX™ not only allowed these students to solve the problem in a way that made sense to them, but the tools and metrics suggested in this and previous chapters allowed us to discover such a novel strategy.

6.3 Conclusions

Our ability to infer why students perform in the way they do seems to have always been important in the U.S. educational system. Early in the 20th century, researchers and educators wanted to test for student intelligence or the ability to function with a high level of proficiency in many, if not all, intellectual pursuits. While studies to measure the tangibles that suggest the presence of the intelligence construct are ongoing, educators and educational researchers have begun showing more interest in measuring student understanding. The present standards suggest we want to know if students can “use their minds well.” We want them to demonstrate that they can explain a concept or topic using logical reasoning and supporting facts. We want them to revise their explanation in light of new facts and we want them to demonstrate an ability apply their explanation in new situations.

While certain forms of pencil and paper tests allow us to assess such explanations, most do not. Even if they did, the data reviewed here suggests that the resulting student evaluations could be confounded by a student’s literacy in the English language. While understanding English is important in an assessment of English proficiency, it may mask our ability to determine how well a student understands subjects like qualitative chemistry. Research suggests that pencil and paper assessment may often exhibit similar biases against various gender, ethnic, and socioeconomic groups.

IMMEX™ computer technology allows teachers and researchers a larger window into student thinking because it is able to record not only what a student answered, but how they arrived at their answer. Furthermore, it can deliver different (but conceptually related) problems to students randomly to allow students to apply their knowledge in new, unlearned situations.

This technology offers us the opportunity to see how a student solves problems and how they adapt to similar, but unlearned, situations in the same domain. Accordingly, it would seem to offer us the chance to begin accurately inferring the presence of constructs like understanding that we feel are important measures of student ability and educational
effectiveness. This is something we have yet to do very well with constructs like intelligence.

The significant correlation between a standard measure of student understanding like the teacher using a rubric to assess a written description and the IMMEX™ measure of the type of information used in and the success rate of solving a problem suggests this technology may offer an important insight into student understanding. Furthermore, the low correlation between the IMMEX™ measure and other content and guessing metrics suggest we are measuring a different dimension of the student here. While familiarity with content is an important part of what we are measuring, it is only one part. IMMEX™ allows us to go beyond merely testing if a student is knowledgeable (the lowest level in Bloom’s taxonomy) to assess something that current tests are missing.

In fact, the results reported in this chapter suggest we can discern how well a student can explain their thinking, use facts to support an explanation, and apply that explanation to situations that are similar to situations they have practiced before. In short, we can infer if a student understands the content we are assessing. Moreover, we can finally begin to address Bloom’s lament that the educational system overvalues and assesses knowledge while ignoring the more important aspects of education. Most importantly, because computer software can make these higher order skills relatively easy to measure accurately, the probability that this technology will improve student thinking and learning seems high.
Chapter 7. Conclusions and future directions

The effectiveness of the education system in the United States has been a topic of debate on and off for more than a century. The debate seems to become particularly intense among policy makers and within the general population at times of crises or when something suggests the graduates of our primary and secondary educational programs don’t “measure up.” In the first part of the last century the military requirements of World War I demanded we identify the most intelligent applicants to facilitate the rapid expansion of the military’s officer corps. In the late 1950’s it was the Soviet threat that suggested we needed improvements in math and science education. In the late 1980’s declining national and international test scores suggested that the educational system was not producing students who could “use their minds well.”

The supposed shortfalls in our system of education have prompted various reform proposals. In the earlier part of the century it was expanding access to higher levels of education. In the 1960’s it was “new math”. More recently, some propose reforms such as a return to basics or more rote memorization and practice of basic skills. Others suggest imposing more stringent accountability systems on schools and educators. Still others advocate the addition of technology to the curriculum. While the focus of each reform varies, they all presuppose that there is a problem with the education system that needs fixing. Each can point to data which allows them to infer that the current system is ineffective and is, therefore, failing.

Ultimately any debate over how well the educational system and the students in that system perform becomes a question of testing and assessment against certain standards. At present, polls in the US suggest that both policy-makers and the general citizenry believe more testing will improve the system. However, the literature suggests that our current testing methods, especially high stakes, multiple choice tests, will not accomplish this goal. At best these systems offer a way to measure how well students can recall pieces of information in a testing situation that bears little resemblance to the context in which the information was learned or in which it will be applied. Moreover, the research suggests that increasing these types of tests will ultimately reduce the curriculum in our schools to teaching students how to do well on tests that have no relationship to the way students will be expected to use what they have learned in the future. Consequently, before proposing changes to the system of education, we need a method to accurately gauge the effectiveness of those changes on student learning. It was my hypothesis that, if applied correctly, modern computer technology affords us the opportunity to improve the accuracy of our assessments of student learning. Furthermore, I suggested that these new methods would offer new insights into student learning and understanding that could affect teaching methods which, in turn, could improve student learning. Here again, these hypotheses were formed using the national and California state standards as a guide to what teachers should teach and students should learn.

During the last fifty years of the twentieth century, a sizable amount of literature discussed how to improve educational and psychological testing. In essence, this body of literature suggests that accurate inferences about student achievement in education will result if a test or assessment:
• evaluates students over the content they have been taught in a format they can recognize.
• consistently evaluates students in both the near- and the long-term against a public, objective, understandable, reliable, and realistic standard.
• accurately evaluates the traits of students we are really interested in.

While some have argued that the new introduction of technology into American classrooms can improve student learning and achievement, I have approached the ongoing debate from another angle. I argue that we should use technology to first improve the accuracy of our assessments in order to have greater insights into student learning and achievement. Once we are able to collect the data we need to make valid inferences about student learning and achievement, we will have a clearer vision about how to improve our system of education. This dissertation offers a “proof” of that concept in one, specific educational domain.

Since accurate assessments are key to the concept I have proposed, I used the validity frameworks as suggested by the literature and as accepted in the field of educational measurement as a guide to my research and as a model for managing the integration of technology into schools. The results suggest that technology can not only improve the validity of the inferences we make about student learning, but, more importantly, that these inferences suggest reforms that will truly improve the learning of those we are trying to educate. I will first review the improvements in the areas of content, criterion, and construct validity and then address the reforms these findings suggest in the section entitled “Future research directions” that follows.

7.1 Improvements in content validity and its implications.

One of the most troubling aspects of our current system of testing is that many of these tests do not match the curriculum students have experienced. The literature suggests that this is true not only for large-scale international tests, but for national and state tests as well. Testing students on knowledge they have never been taught and then implying that these students are less able or are lower achievers is not a justifiable conclusion. The conclusion that schools are ineffective based on this data may, likewise, be unjustified. Among others, two aspects of standardized, large-scale tests seem to make them especially vulnerable to this type of invalidity. First, because they must often test a large domain using a minimum of questions, they are often secret. Second, in order to validate them in their present form, individual teachers cannot adapt them to a specific classroom context or instructional curriculum. Therefore, if the test questions are multiple choice, but the students learned in a more holistic context, few options are available to the teacher. The research suggests that, depending on the importance of the test, the teacher will tailor the curriculum to match the test. If the test is multiple choice or another type of selected response test, the result is often a curriculum that teaches students to memorize facts in order to score well on the test. Here again, the literature suggests that students are unlikely to retain or be capable of applying those facts to real world problems.
This suggests that more local assessments would be the least likely to suffer from such mismatches between test and instructional content; however, this has been difficult to validate. Furthermore, comparing the content of a test in one classroom with that of another is difficult both because of number of classrooms and the way questions may be posed.

The adaptive nature of technology and its ability to record large amounts of disaggregated data seem to offer us another alternative to the current system. Software that allows the teacher to create or adapt problems to his or her actual curriculum, makes such problems publicly available, and allows teachers to assess students in a way that is sensitive to the instructional context they experienced is available. These abilities were demonstrated using the Interactive Multi-media Exercises (IMMEX™) software.

When applied in the subject area of qualitative chemistry, IMMEX™ data allowed us to determine that there was a high correlation between what was tested and what was taught. In fact, it made such a determination relatively easy. Since IMMEX™ problems are available for review in the public domain, policy-makers, experts, teachers, students, etc. have the opportunity to review them before, during and after use. In addition, modern test theory allows the evaluation of actual student IMMEX™ performances to quantify the match between test and curriculum content. The student performances in this research suggest that students had been exposed to the content evaluated by each Hazmat clone, and that each clone measured different aspects of the curriculum. Since the literature suggests that a relationship exists between test and curricular content regardless of subject area, the IMMEX™ problems developed for other domains should allow similar evaluations in those domains as well.

Consequently, while it is virtually impossible to determine the validity of many local assessments at present, the data generated by IMMEX™ allows us a way to at least begin such an investigation without incurring large expenditures in time or money.

7.2 Improvements in criterion validity and its implications

In order to improve teaching and learning, both teachers and students must know what the objective of the curriculum is and how close each student is to achieving that objective. Unfortunately, many of our current tests measure against different standards. Probably the most notable example involve state tests that measure student ability against one standard and the National Assessment of Educational Progress (NAEP) that measures against a very different standard. Consequently, many of these state tests rate the proficiency of their students high, while the NAEP rates the proficiency of a statistical sample of the same students as very low. Unless one reconciles the standards against which each test measures, the results cannot be compared. Furthermore, many of the large scale tests have reliabilities that make their ability to predict future student performance questionable. Often their ability to predict future performance is further reduced by variables such as gender, ethnic, or socioeconomic status.
In this study, the ability of early problem solving to predict future problem solving performances was also low when the performance was classified dichotomously as a correct or incorrect solution. However, when we considered how the student solved the problem, earlier performances did predict how students would attempt to solve problems in the future.

One of the problems with using an earlier performance to predict a later one is the likelihood that students will perform the same task in different ways on different occasions. Heretofore, this has been a major conundrum for educational measurement. In fact, the methodology used in this study added to such a conundrum as different clones were presented to different students in random order. This almost assured that the students would have to use different strategies to solve problems over time.

Analysis of the data in this study suggested a solution to this conundrum. The odds a certain strategy actually produced a correct solution to the problem and the amount of data used in the strategy, suggested three obvious strategy types. Students using "limited" strategies investigate very little information before attempting to solve a problem. Conversely, in some strategies the students viewed more than enough information to solve the problem. These strategies were termed "prolific" for this reason. Both of these strategy types have low odds of solving the problem in Ms. Lindsay's classes. On the other hand, students who focused almost exclusively on the information that was germane to identifying the unknown, had better than even odds of solving the problem. These strategies were grouped into the "efficient" strategy type. By focusing on strategy types rather than specific strategies, we can determine if a student used successful strategies to solve problems rather than if a student used a particular strategy to solve problems.

More importantly, the research I have reported here suggests that the first strategy type a student uses on a problem will likely indicate the type of strategy that student will use on the remaining clones of that problem. Furthermore, the student's first strategy is likely to indicate the type of strategy a student will use on other similar problems in the future. While this demonstrates the criterion validity of the IMMEX\textsuperscript{TM} computer technology, I believe it has even greater importance in the classroom. These findings suggest that this technology can be an important tool in the classroom. A teacher can begin to identify students who are struggling with problems almost from the time a student attempts to solve his or her first one. Furthermore, the data suggests that students who use limited strategies may need help learning content as their grades tend to be lower than students using the other strategy types. On the other hand, students using prolific strategies may need help with applying content. As such, these general classifications can help an educator develop and tailor pedagogy to the needs of specific students.

The "hands-on," authentic assessment demonstrated in this research evaluate both the ability of the student to learn content as well as the ability of the student to apply that content. As such, it not only seems to fulfill the criteria advocated by the national standards, it also allows teachers to evaluate students against standards that heavily favor knowing content. The California state standards are just one example of such standards.
While this tool might not demonstrate the same effectiveness using other standards, given the current national and state expectations, it functions very effectively. Moreover, unlike many of the national and state tests currently in use, no gender or socioeconomic correlations, which suggest possible bias, were noted in the IMMEX™ data. I did note, however, that while non-white students tended to favor limited strategy types, white students tended to disproportionately use prolific strategies. There were no significant ethnic differences among the students using efficient strategies.

7.3 Improvements in construct validity and its implications
The national and state standards, as well as much of the available literature suggest that, ultimately, policy-makers, educators and other stakeholders in education want to know why a student can do what s/he does. Until the latter half of the twentieth century, this why was most often explained by the psychological trait of intelligence. More recently, the educational community has desired to know, not what native ability produces an observed result, but how well the student can explain how or why they arrived at a result. This is the construct called understanding. As a psychological construct, understanding cannot be measured directly but must be inferred from other behaviors or objects produced by the student.

For the purposes of this research, two metrics of understanding were employed. First, the teacher in the class developed a rubric to assess the logic her students used and noted when solving Hazmat problems during an assessment. Second, I developed a metric to determine if a student had demonstrated understanding based on his or her IMMEX™ performance. In essence, if a student had enough information to solve a problem and actually solved the problem, their performance was classified as understanding. As expected, the correlation between these two metrics was very large and significant; furthermore, the two metrics had much lower correlation coefficients with indicators that were indicative of other constructs like guessing, and basic content knowledge. These results suggest that we can infer when students demonstrate understanding from their actual performances on these types of authentic problems. Most pencil and paper forms of assessment do not allow such an insight, and those that do usually make it incumbent on the student to decide what s/he will and will not write down for an assessor to see. IMMEX™ records everything the student does and so provides a more accurate picture of the student’s performance.

With the technology used here, we apparently have a much larger and more accurate window into student thinking because we see not only how a student answered, but how s/he arrived at that answer. As such we can begin to assess the much deeper levels of learning suggested in Bloom’s taxonomy and often called the “Higher Order Thinking Skills” today. If as the literature suggests, “what gets tested, gets taught,” the ability to easily and accurately assess these skills against our present standards could have a dramatic effect on education.

In the first chapter of this dissertation, I indicated that I wanted to explore the question of how modern technology could be used to offer more accurate insights into student
learning and understanding. I hypothesized that the technology, if properly employed, could improve the accuracy of our insights and provide new and useful insights into how students learn and understand. Since the data produced by most of our present testing methods were being used in a way they were not intended, I used the validity standards in the APA’s *Standards of Psychological and Educational Testing* as a technology management framework in this study. I believe the results in each area not only allow one to conclude that this use of computer technology improved the accuracy of our inferences, but that this research yielded important new insights into student learning and understanding. I also believe that these results could affect new educational models. The final section in this dissertation will suggest future research inquiries.

A caveat must be attached to the findings suggested by the research in this dissertation. As pointed about above, this research was done in light of the present national and California state standards. As such, the tool was designed to gather data about how well students knew and could apply knowledge in a particular domain. Should the standards change, IMMEX™ might not be the most appropriate tool to use in evaluating student performance against that new standard. For example, if national or state standards required only that students memorize a set of facts in order to be judged proficient learners in a particular domain, the IMMEX™ technology might be judged to be too sophisticated for determining if a student had met the standard.

In a similar manner, as presently constituted, IMMEX™ seems inappropriate for assessing some types of student cognitive abilities. IMMEX™ problems focus on the ability of students to reason deductively from a problem presentation to its solution, using data elimination and reduction. Current IMMEX™ problems, like chemistry or genetics problems, require students to develop and examine different possibilities, then eliminate those that cannot be true. More inductive or creative processes, such as painting or writing poetry, are probably not the types of problems that could be modeled using the IMMEX™ tool.

Future modifications to the IMMEX™ authoring and presentation modules will allow students to examine information and then compose textual input based on their reaction to that information. Educators or other evaluators will still have to assess these textual inputs. While the tools demonstrated here will allow for a more automated analysis of how the students transited a problem space, assessing the meaning of what Krathwohl calls the “affective domain” (Krathwohl, Bloom et al. 1964) seems to fall outside the technological limitations of IMMEX™.

### 7.4 Technology management and policy implications

Because technology has become a part of virtually every aspect of our lives, many have assumed that by merely adding modern computer technology to our classrooms, American education will improve. We have wagered hundreds of millions of dollars that such an assumption is true. While many are willing to accept the truth of such an assumption with blind faith, others have attempted to prove or disprove it. These attempts have met with varying degrees of success. Like other assessments in education,
assessments of success or progress depend on what is measured and how one evaluates the results of such a measurement. Sound educational policy at all levels depends on accurate inferences from these assessments. As noted previously, Means argues that if we cannot demonstrate the effectiveness of integrating technology into our schools, the perpetrators of this latest educational "fad" are likely to be severely castigated. (Means and Olson 1997)

In order to determine whether modern computer technology will improve our system of education in the United States, we must first determine what we want technology to do for education. If we expect it to accomplish the same thing we currently expect from pencil and paper, or other less expensive and less computationally powerful tools like a handheld calculator, the monetary costs seem to outweigh the benefits dramatically. If the time required to learn, use or integrate the technology into a classroom adversely affects what is currently taught and learned in such a classroom, current experience suggests the technology will not be used. If the technology is not reliable, teachers are likely to cast it aside. To validate the assumption that it works, we must demonstrate that technology has the ability to provide a benefit which is not reasonably achievable with the tools currently at our disposal. Only then can we move on to the policy questions of how to integrate it effectively and on a large scale.

Depending on their format and content, current methods of assessing student knowledge seem adequate for many policy makers. In fact, the calls for dramatic increases in this same type of testing are currently widespread. Perhaps those voicing such demands are of the opinion that we must ensure that students can merely pick the correct answer from a list of other deceptive, but wrong answers, or that students can apply a memorized algorithm in a particular situation. Maybe proponents of such testing are unaware that many of the most prevalent large-scale tests cause the curriculum in schools to teach students to memorize rather than understand concepts or that test scores often increase merely because students and teachers learn the test. Perhaps they are unconcerned that certain assessment formats penalize some groups of students more than others do. On the other hand, perhaps calls for more of the same testing arise because the supporters of current testing systems are unaware that technology offers us the opportunity to assess student learning more deeply and more accurately than ever before. Somewhat paradoxically, using technology for such assessments also allows us to determine whether other educational applications of technology are effective in improving what students learn and understand.

If the various national and state educational standards are to be believed, what the American public really wants from its schools are knowledgeable students who can "use their minds well." They expect students who can demonstrate that they understand concepts by applying these concepts in ways and in situations that others may have never anticipated. Merely training students to accomplish a task, even if we hope they can adapt such training and apply it to another dissimilar task, may have under served us in the past; and given the rapidly changing world and dramatic growth of knowledge, it is unlikely to serve us well in the future. My research suggests that technology, if it is
employed properly, offers us new and previously unavailable opportunities in educational assessment under our current set of national and the California standards.

Modern computer technology affords us new options in education and, more particularly, for assessing the student learning and understanding required by current standards. However, these options will achieve their promise only if we are willing to use them appropriately. The static nature of current pencil and paper tests is a major detriment in our efforts to make accurate inferences about student learning and understanding in light of current standards. Evaluating students against static learning models has provided limited information that many have used to arrive at conclusions the data was never intended to support. In general, three types of inaccuracy account for such faulty conclusions: content inaccuracy, criterion inaccuracy, and construct inaccuracy.

Because large-scale tests are designed to be administered to students from diverse schools and taught by different teachers, test makers can only theorize what concepts students have actually been taught. The literature makes clear that, in fact, test content most often does not match what these students have actually studied. Furthermore, the literature suggests that students who have studied the content being tested score better on such a test than those students who have not studied the same content. For this reason, the results of such tests seem to support conclusions about how well aligned a test is with a curriculum, not about how effective a student is at learning or a teacher is at teaching a specific curriculum. Test developers must likewise consider the context in which a teacher delivered, or a student has studied, the content when developing test formats. Consequently, it seems that technology must allow us to adapt test content to what is studied and taught in a particular course.

The criteria used to evaluate a student’s performance on a test also vary dramatically between national, state, districts and schools, and even between classrooms in the same school. The very nature of the U.S. educational system virtually guarantees that numerous and diverse standards will continue to exist in the future. Unfortunately, pencil and paper tests most often allow only one of these entities to evaluate a student’s performance against a single standard. The large cost of duplicating student work, the secrecy of test forms, and the propensity to make evaluations more objective by assessing only the correctness of an answer all contribute to the inaccuracy of results. In most instances, evaluators report only the results of their evaluations of student performance. They seldom include the standards used to make such evaluations. Consequently, to improve the status quo, technology must allow us to share disaggregated student performance data and allow various evaluators to judge a student’s performance against a standard appropriate to the specific need. It must also allow users to tie a standard to an evaluation so that users can weigh both in inferring student ability.

Finally, current pencil and paper tests seem ill suited to provide the data necessary to accurately infer whether a student possesses traits like understanding. A century of intelligence testing should have convinced us of this. While we can accurately discern these traits in students, our present method of actual student observation is both costly and open to observer bias. Furthermore, requiring a student to duplicate a specific model
of problem solving, or only evaluating a student’s answer to a problem, provides little
data with which to make accurate inferences about intangible traits like understanding.
Consequently, when the standards require students to demonstrate such an ability,
technology must allow students to explore a problem space and explain their solution to a
problem in a way that is meaningful to that student. It should allow us the ability to
observe a student performance by recording how a student arrived at a solution, not only
what solution the student reached.

The foregoing suggests we consider some essential factors when deciding which
computer technologies to integrate into our educational system, especially for the
purposes of assessing student learning and understanding. The research and findings in
previous chapters suggest that to make the type of accurate inferences we currently
demand about student learning, technology must:

- allow teachers to adapt the content of an assessment to specific teaching and learning
  situations easily.
- record actual student performance data and allow appropriate evaluators access to
  both that raw performance data and the actual task or problem that the student was
  working on when performing in that manner.
- function as a “tool” in the hands of the student, allowing the student to solve a
  problem or accomplish a task in a manner that makes sense to the student.
  Accordingly, the technology should not evaluate students against a particular model
  or constrain students to act in a certain manner.

These standards for integrating educational technology not only follow the Standards for
Educational and Psychological Testing, but they also closely match Dede’s admonition
that schools systems integrate technology from the “top down, bottom up and middle
out.” (Dede 1998) More specifically, these integration guidelines allow stakeholders in
the educational enterprise to accurately evaluate student performances and exercise their
responsibility for holding schools accountable. These stakeholders include parents
evaluating the work of their own children, and state policy makers, district officials and
local administrators evaluating the adequacy of state and local schools. Moreover, it
allows stakeholders to apply various standards to student performances and to publish
both the standards and the evaluation of the performances against those standards jointly.
The guidelines suggested here also give teachers in the “middle” of the system the ability
to adapt assessments and curricular content to the needs of their students, instead of
constraining their curriculum to a specific test. Finally, the guidelines allow students to
demonstrate an ability to solve problems in a way that makes sense to them, instead of
memorizing content or algorithms that they must voice back to their teacher in a rote
manner. In short, given our current standards, this path of technology integration into our
system of educational assessment allows for more accurate and timely assessments at all
levels of the educational endeavor.

This application of technology could also have profound policy implications. We spend
billions of dollars annually on education at the national, state and local levels. Policy
makers predicate these large outlays on the assumption that we can accurately measure
student learning and take appropriate corrective action when our measurements indicate a lack of student achievement. At best, our current tests provide a very limited picture of what students know and what they can do with that knowledge even though the standards suggest students must do both. These tests are often biased against certain groups and encourage curricula that teach memorization of facts to pass a specific test. Nevertheless, the major focus of educational policy at the national and state level either relies on the results of these tests or seeks to increase the number of such tests in the nation’s schools. As such, current policy risks further constraining the curriculum in our schools toward rote memorization and the development of “fads” are unlikely to accomplish the objectives that the current standards suggest we expect from our educational system. A technology integration plan focusing first on gathering the data that will allow us to make valid inferences about student learning and understanding, like the one proposed here, is one way to mitigate these risks.

In light of the present standards, managing technology in this way will also provide the means necessary to both suggest more appropriate educational policy options and to evaluate those options with greater accuracy in the future. For example, developing assessment methods unencumbered by racial or socioeconomic biases could dramatically effect how Title I funds are allocated by the Congress and used by the states. Moreover, since underrepresented minorities and the poor are most likely to be taught rote memorization to do well on current tests, a change in testing policy is likely to have a positive effect in encouraging curriculum that truly educates these students. Nor is the private sector exempt from the benefits of such policy changes. New assessment methods could have a profound effect on which candidates colleges and universities admit and whom employers choose to hire. Most importantly, because the research suggests that the methods advocated in previous chapters will identify important intervention points, decision-makers can target policies to address specific shortcomings in the current system of education more directly. For example, a policy of reducing class size might be inappropriate if the real cause of a student’s inability to achieve is related to curricular content or an inability to apply that content instead of how much time a teacher can spend with each student.

The literature cited in preceding chapters suggests that the current system of assessing student achievement in the United States is too limited to provide the data we need to make the inferences required by current standards. Consequently, we are unlikely to know if the present educational system is functioning as expected, or if it requires change. We are also unlikely to know which policy changes will affect a more beneficial system. However, the research described in previous chapters also suggests a way to manage technology to improve both education and education policy in the United States.

7.5 Future research directions
While the results in this dissertation appear significant for and apparently generalize to a certain part of the high school population in America, it would be prudent to repeat this entire study in an area where both the ethnic and socioeconomic diversity of the students
is larger or is composed of more non-white students and/or students from lower socioeconomic backgrounds. Although non-white students were present in this study, they were outnumbered by white students almost 2 to 1. There was, moreover, only one African American student in this sample. Also, as most of the students in this study lived in an upper middle-class suburban setting, a more urban setting would be advisable if one wanted to extrapolate these findings to a more general population. Finally, because the teacher in this study was a dedicated, conscientious professional, it might be informative to conduct this test in the classroom of one or more teachers that did not meet these high standards.

Because this study was testing for concurrent and predictive criterion validity, the teacher in this study did nothing to overtly correct the deficiencies in student performances between IMMEX™ problems. A follow-up study that made early interventions with the students demonstrating less than effective strategies might allow researchers to suggest and test important interventions to improve student learning and understanding. Possible interventions include:

- individual students review search path maps and suggest steps to omit or to add that would improve a specific performance.
- students, individually or as a group, review search path maps of limited, efficient, and proficient strategies and suggest what makes them effective or ineffective.
- students review the performances of others in dyads or small groups to suggest what makes them effective or ineffective.
- students work in groups of two to actually perform IMMEX™ problems. The literature suggests that heterogeneous ability groups would be best.
- require students to perform enough Hazmat clones so that students attempt all possible combinations before being assessed.

One of the most important findings suggested by the research in the previous chapters is that we can almost immediately identify two groups of students that will have difficulty solving qualitative chemistry problems. Furthermore, students who used limited types of strategies apparently have lower grades than students using prolific strategies. This suggests that the most effective intervention for a teacher to make in each group might be dramatically different. Specifically, students using limited strategies might benefit from learning more content, whereas students using prolific strategies might benefit more from meta-cognitive training, learning problem solving skills, or studying how to apply the content they seem to have already learned but cannot apply.

On going research suggests that these same patterns occur in other high school science courses such as biology. Obviously, conducting studies similar to the one reported here in other aspects of the chemistry curriculum and in other science courses seems warranted. Moreover, because IMMEX™ problems have been designed by teachers of English, history, mathematics and the language arts, studying student cognition in these domains might be a rich source for developing other models of learning.
Elementary and middle schools are also logical candidates for similar studies. While a large number of studies are planned for college undergraduate and graduate schools, I know of no plans to investigate the learning and understanding of primary school students in the manner described. A study of this sort seems not only a logical way to assess learning in the early years of student development, but also a way to assess changes in student learning from primary, through secondary, and into collegiate years. While an enormous undertaking, the valuable insights into how humans learn and the most effective ways to help that learning seem well worth the effort.

Finally, it seems that college and university admissions programs could benefit from using a system such as IMMEX™. Repeatedly during this research and during my years as a high school teacher, professors lamented that even their students with the highest high school grade point averages did not understand basic concepts in a particular domain such as calculus or chemistry. Moreover, these instructors felt most of their students possessed little ability to actually apply the content of a particular domain to a real world problem or to explain a basic concept from that domain. Given the low predictive value of present standardized tests, the results in previous chapters suggest that evaluating students solving actual problems using a tool like IMMEX™ would allow a low-cost and expeditious way to assess student understanding. It would also do so in a manner that undergraduate and graduate schools could evaluate against standards they felt predicted success at their school.

Although there are many other interventions that can be attempted, those listed above seem an obvious way to begin exploiting the capabilities of the type of technology I have described. Applying technology in this way will also cause students and teachers to reflect on student thinking and will allow more accurate inferences about what students know and can do. Moreover, the immediate application of this type of technology to American education seems particularly timely. Policy makers and other stakeholders are demanding an accountable education system that produces improvements in student learning and achievement. In light of these requirements and our current national and state standards, however, experience suggests that merely increasing the frequency of the current mode of testing will provide little new information and will not encourage students to “use their minds well.” In fact, history suggests it may have the opposite effect. The literature suggests that one way to affect an outstanding education system are valid assessments designed to yield the insights into student learning and understanding we require. The benefits of integrating IMMEX™-like technology into our schools, as demonstrated in this dissertation, are not only more accurate inferences of what students have learned, but new insights that will allow all stakeholders in the U.S. educational system to work together to provide each and every student the opportunity to learn as much as possible.
Appendices
Appendix 1 – Initial Student Survey
IMMEX Labs / The Massachusetts Institute of Technology

Student Survey

This is a confidential questionnaire. Your individual answers will not be shared with your teacher and will be held in confidence by the researcher. Please answer each question as honestly and accurately as you can. Thanks!

Student IMMEX ID: ________________________________

1. What is your birth date? (mm/dd/yy) _____/_____/_____

2. What is your gender? □ Female □ Male

3. What is your ethnicity:
   □ African-American □ Asian-American □ Latino / Latina
   □ Multi-ethnic/Biracial □ Native-American □ White
   □ Other: ________________________________ □ Prefer not to state

   Do you participate in:

4. Sports or sport related teams? □ Yes □ No

5. Music, band, or drama? □ Yes □ No

6. Academic extra-curricular activities (including speech and debate)? □ Yes □ No

7. Self-development or community service activities outside of school? □ Yes □ No

8. If you have a job, how many hours do you work, on average, each week (including weekends)? ____________ hours

9. Do you have easy access to a car when you want to drive? □ Yes □ No

10. What are your plans after high school? (Please choose one)
     □ Work / Trade School □ 2 year community college □ 4 year college / university
     □ Military Service □ Undecided □ Other: ________________

11. What is your zip code? __________________________

12. What is your approximate, current high school GPA? ________________
13. Put a check next to the following statement that best describes the parent you live with or the parent with the most education?

- Never completed high school
- Completed high school, but has no college experience
- Some college experience, but has no college degree
- At least one bachelor's degree
- At least one Master's
- At least one Doctorate (Medical, Lawyer, Ph.D., etc.)

14. What was your first and second semester grade in each of the following classes: (if you didn't take a class, please write N/A in the space provided):

<table>
<thead>
<tr>
<th>Subject</th>
<th>First Semester</th>
<th>Second Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algebra 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algebra 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

15. How many AP or honors classes have you taken (all those you've completed or are enrolled in now)?

16. What are your approximate grades in Chemistry?

Please answer questions 17 – 19 using the scale provided:

<table>
<thead>
<tr>
<th>Question</th>
<th>Not at all</th>
<th>Very little</th>
<th>Some</th>
<th>Quite a bit</th>
<th>A lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>17. How well do you like this class?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>18. How well do you like this teacher</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>19. How well do you think this teacher likes you?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

20. Have you ever tutored anyone else in Chemistry?  □ Yes  □ No

21. On average, how much time do you spend actually working on Chemistry (reading, solving homework problems, etc) each night?

___________ minutes
22. On average, how often do you speak up in class? (Choose one)
   □ Never
   □ Less than once a week
   □ Once or twice a week
   □ More than twice a week

23. Do you generally “run out of time” when you take tests in Chemistry?
   □ Yes □ No

24. During non-lab times, where do you sit in your classroom:
   □ Front of the room
   □ Middle of the room
   □ Back of the room

For questions 25 – 35 please check True or False

<table>
<thead>
<tr>
<th>Question</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>25. The information we are learning in this class will be important for me to know in the future.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26. I seek help from my teacher outside of class (during break, before school or after school) when I have questions about Chemistry.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27. My teacher’s evaluation of what I know about Chemistry is accurate.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Please explain briefly:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28. I have a clear idea of what the teacher requires of me to do well in this class.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29. The requirements of what it takes to do well in this class are realistic and not only based on what this teacher feels is important.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30. I enjoy doing “labs” in Chemistry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31. Learning Chemistry is frustrating.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32. In general, I don’t get a new concept (for example, chemical bonding) in Chemistry until I think about it for at least a day.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
33. In general, I don’t get a new concept (for example, “precipitation or ppt”) until I use it in a “lab”.

☐ True  ☐ False

34. In general, I don’t get a new concept (for example, “stoichiometry”) until I apply it in homework.

☐ True  ☐ False

35. I can memorize things for this class, take a test, then forget about them and still get a passing grade in this class.

☐ True  ☐ False

In general, how often do you do the following when solving problems:

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Often</th>
<th>Very Often</th>
</tr>
</thead>
<tbody>
<tr>
<td>36. Look at the available information before making a final decision?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>37. Pick the first solution that comes to mind?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>38. Pick the solution that seems easiest?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>39. Try to look for more than one solution to the problem?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>40. Look for information that will eliminate a solution?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>41. Look for information that will support a solution?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>42. Ignore information that goes against what you think the solution is?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>43. Only use information that goes along with what you think the solution is?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>44. Talk to your teacher or other people to get information?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>45. Routinely ask yourself what information is important?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

46. Have you used the Interactive Multi-Media Exercises (IMMEX) software to solve problems before? (If you don’t know what IMMEX is, please select NO).

☐ Yes  ☐ No

Thanks for taking the time to answer this survey accurately!
Appendix 2 – Initial Teacher Survey

1. Predict the class rank of this student in the qualitative chemistry unit.
2. How often has this student’s parent initiated contact with you during the first semester?
3. How would you rate the appearance of this student (unkempt, average, or well dressed)?
4. In your opinion, is this student motivated to do well?
5. Is this student disruptive in class?
6. What is your overall assessment (grade or similar) of this student?
7. In your opinion, how hard does this student work in your class (above potential, at potential, or below potential)?
8. Do you feel this student is inquisitive?
9. Do you enjoy having this student in your class?
10. On average, how often does this student speak up in class each week?
Appendix 3 – Hazmat Student worksheet
HAZMAT, holiday special

Name: __________________________ Period: __________

Fill in each section for points towards your grade

Flame Test Description/positive ion

List of negative ions:
- OH⁻, CO₃²⁻, SO₄²⁻, NO₃⁻, Cl⁻

Physical Tests

Chemical Tests
Remember that tests with ionic compounds (K₂CO₃, Ba(NO₃)₂, Na₂SO₄, NaOH, and HCl) give you a net ionic equation to figure out!

Appendix D  Table of Solubilities of Some Ionic Compounds in Water

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Any negative ion</td>
<td>+</td>
<td>Alkali metal ions: (Li⁺, Na⁺, K⁺, Rb⁺, or Cs⁺)</td>
</tr>
<tr>
<td>Any negative ion</td>
<td>+</td>
<td>Ammonium ion, NH₄⁺</td>
</tr>
<tr>
<td>Nitrate, NO₃⁻</td>
<td>+</td>
<td>Any positive ion</td>
</tr>
<tr>
<td>Acetate, CH₃COO⁻</td>
<td>+</td>
<td>Any positive ion except Ag⁺ or Hg⁺</td>
</tr>
<tr>
<td>Chlorine, Cl⁻, or Bromide, Br⁻, or Iodide, I⁻</td>
<td>+</td>
<td>Ag⁺, Pb⁺, Hg⁺, or Cu⁺</td>
</tr>
<tr>
<td>Any other positive ion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulfate, SO₄²⁻</td>
<td>+</td>
<td>Ca⁺⁺, Sr⁺⁺, Ba⁺⁺, Ra⁺⁺, Ag⁺⁺, or Pb⁺⁺</td>
</tr>
<tr>
<td>Any other positive ion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulfide, S²⁻</td>
<td>+</td>
<td>Alkali ions or NH₄⁺, Be²⁺, Mg²⁺, Ca²⁺, Sr²⁺, Ba²⁺, or Ra²⁺</td>
</tr>
<tr>
<td>Any other positive ion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydroxide, OH⁻</td>
<td>+</td>
<td>Alkali ions or NH₄⁺</td>
</tr>
<tr>
<td>Any other positive ion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphate, PO₄³⁻, or Carbonate, CO₃²⁻, or Sulfite, SO₃²⁻</td>
<td>+</td>
<td>Alkali ions or NH₄⁺</td>
</tr>
<tr>
<td>Any other positive ion</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

155
Appendix 4 – Unit 12 pencil and paper test
DO NOT WRITE ON THIS TEST! USE IT AS A COVER SHEET AND WRITE ON THE ANSWER SHEET.

CALCULATE AND CONVERT TO SCIENTIFIC NOTATION:

1. \(3.15 \times 10^3\)  \(2.50 \times 10^{-7}\)

2. \(4.8 \times 10^{17}\)

\(2 \times 10^9\) \(6.0 \times 10^{-11}\)

4. \(4.0 \times 7.5\)

SOLVE THE FOLLOWING PROBLEMS. ADJUST ANSWERS USING THE PROPER NUMBER OF SIGNIFICANT FIGURES. (Convert answers to scientific notation):

3. \((7.00) (9.0 \times 10^{-2})\)

5. Magnesium chloride

6. Calcium phosphate

7. Sulfur trioxide

WRITE THE FORMULAS FOR THE FOLLOWING MOLECULES. ADJUST FOR PROPER VALENCE.

8. BaCO₃

9. CaSO₄

10. SiF₄

BALANCE THE FOLLOWING EQUATIONS. I.D. #11, 12 (N) Endothermic (X) Exothermic

I.D. #13-14 A. (C) Synthesis (G) Single Replacement (H) Combustion (J) Decomposition (K) Double Replacement

11. \(\text{Al}_2\text{O}_3(s) \rightarrow \text{Al}(s) + \text{O}_2\) \(3351.4 \text{ kJ} = \Delta H^0\)

12. \(\text{Sn}(s) + \text{Cl}_2(g) \rightarrow \text{SnCl}_4\) \(-511.3 \text{ kJ} = \Delta H^0\)

13. \(\text{C}_2\text{H}_2\text{O} + \text{O}_2 \rightarrow \text{CO}_2 + \text{H}_2\text{O}\)

14. \(\text{Br}_2 + \text{NaI} \rightarrow \text{I}_2 + \text{NaBr}\)

CHOOSE THE LETTER OF WHICH FACTOR OR BRIDGE WILL CONVERT THE FOLLOWING.

15. 2.3 mol of Si\(\rightarrow\) # of Si

J. \(6.02 \times 10^{23}\) K. \(1\) mole L. \# mol Cu M. 22.4 liters

16. 1.4 mol KCl\(\rightarrow\) # of molecules KCl

1 mole \(6.02 \times 10^{23}\) # mol Fe 1 mole

17. 2.5 g MgSO₄\(\rightarrow\) # mol MgSO₄ N. molar mass (g) O. 1 mole P. # mol Fe Q. 1 mole

18. \(7 \times 10^{25}\) CO₂ molecules\(\rightarrow\) # mol CO₂ 1 mole molar mass (g) # mol Cu 22.4 liters

MATCHING. WRITE THE LETTER OF THE TERM THAT BEST FITS EACH DEFINITION.

19. atoms in compounds tend to have the electron configuration of a noble gas, 2 or 8 valence electrons

K. POLAR COVALENT BOND

20. bond with unequal sharing of electrons

L. IONIC BOND

21. bond with equal sharing of electrons

M. DISPERSION FORCES

22. interactions caused by the motion of electrons

N. METALLIC BONDS

O. ELECTRONEGATIVITY

23. bond formed by the electrostatic attraction between a cation (positive ion) and an anion (negative ion)

P. van der Waals forces

Q. NONPOLAR COVALENT BOND

24. molecule in which one end is slightly more negative

R. MALLEABLE

25. the weakest attractions between molecules

S. VSEPR

26. two or more electron dot drawings for same molecule

T. DUCTILE

27. the electrons in the outer energy level by group #

U. HYDROGEN BONDING

28. positive ions surrounded by freely moving electrons

V. VALENCE ELECTRONS

29. the strongest attractions between molecules ex. water

W. OCTET RULE

30. number measures an atom’s attraction to electrons

X. RESONANCE STRUCTURES

31. valence shell electron pair repulsion

Y. POLAR MOLECULE

32. property of metals, can be hammered into shapes

DRAW ELECTRON DOT STRUCTURES FOR THE FOLLOWING

Atoms:

33. property of metals, can be drawn into wire

34. Calcium 35. Oxygen 36. NH₄⁺ 37. CF₄ 38. Cl₂

MULTIPLE CHOICE. CHOOSE THE ANSWER THAT BEST FITS. #39-55

39. The main use for copper is as a(n)

a) alloy in airplane parts b) electrical conductor c) component in jewelry d) magnets

40. Properties of metals result from metallic bonding, in which positively charged cores of metal atoms are surrounded by

a) neutrons b) protons c) tightly bound electrons d) a sea of mobile, valence electrons

41. A molecule with polar bonds can be nonpolar if

a) the bonds are ionic b) the difference in electronegativity is zero c) the polar bonds are oriented so that the “pull” of each polar bond is balanced d) the molecule contains uncharged bonds

42. The melting and boiling points of most molecular or covalently bonded molecules are

a) lower than ionic compounds b) about the same as ionic compounds c) higher than those of most ionic compounds d) sometimes higher, and sometimes lower than those of ionic compounds

43. A covalently bonded molecule

a) forms between a metal and a nonmetal b) forms when electrons are transferred c) forms between an anion and a cation d) when atoms share electrons

44. Ionic compounds

a) dissociate easily in water b) conduct electricity when dissolved in water c) are solid at room temperature d) all of these

157
Which of the molecules shown at the right is polar?

a. A  c. C
b. B  d. D

USE THE CHART FOR THE NEXT SET OF QUESTIONS:

<table>
<thead>
<tr>
<th>ELEMENT</th>
<th>ELECTRONEGATIVITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorine</td>
<td>3.0</td>
</tr>
<tr>
<td>Fluorine</td>
<td>4.0</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>2.1</td>
</tr>
<tr>
<td>Lithium</td>
<td>1.0</td>
</tr>
<tr>
<td>Magnesium</td>
<td>1.2</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>3.0</td>
</tr>
<tr>
<td>Oxygen</td>
<td>3.5</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.8</td>
</tr>
<tr>
<td>Sulfur</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Indicate whether the following bonds are (A) polar covalent (B) nonpolar covalent (C) ionic

46. H-Cl
47. Mg-O
48. K-F
49. H-H
50. N-O
51. S-O

For each of the following, choose the letter of the drawing below that best represents its shape.

(a) Linear
(b) Trigonal planar
(c) Tetrahedral
(d) Trigonal pyramidal
(e) Bent

52. H₂O
53. NH₃
54. BCl₃
55. CF₄
<table>
<thead>
<tr>
<th>Prefix</th>
<th>Number</th>
<th>Prefix</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>mono-</td>
<td>1</td>
<td>penta-</td>
<td>5</td>
</tr>
<tr>
<td>di-</td>
<td>2</td>
<td>hexa-</td>
<td>6</td>
</tr>
<tr>
<td>tri-</td>
<td>3</td>
<td>hepta-</td>
<td>7</td>
</tr>
<tr>
<td>tetra-</td>
<td>4</td>
<td>octa-</td>
<td>8</td>
</tr>
</tbody>
</table>

**TRANSITION METALS**

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>He</td>
<td>Ne</td>
<td>20</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
</tr>
</tbody>
</table>

**INNER TRANSITION METALS**

<table>
<thead>
<tr>
<th>34</th>
<th>35</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>47</td>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
<td>57</td>
</tr>
<tr>
<td>58</td>
<td>59</td>
<td>60</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
<td>65</td>
<td>66</td>
<td>67</td>
<td>68</td>
<td>69</td>
</tr>
<tr>
<td>70</td>
<td>71</td>
<td>72</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>76</td>
<td>77</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>81</td>
</tr>
<tr>
<td>82</td>
<td>83</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>94</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>104</td>
<td>105</td>
</tr>
<tr>
<td>106</td>
<td>107</td>
<td>108</td>
<td>109</td>
<td>110</td>
<td>111</td>
<td>112</td>
<td>113</td>
<td>114</td>
<td>115</td>
<td>116</td>
<td>117</td>
</tr>
<tr>
<td>118</td>
<td>119</td>
<td>120</td>
<td>121</td>
<td>122</td>
<td>123</td>
<td>124</td>
<td>125</td>
<td>126</td>
<td>127</td>
<td>128</td>
<td>129</td>
</tr>
<tr>
<td>130</td>
<td>131</td>
<td>132</td>
<td>133</td>
<td>134</td>
<td>135</td>
<td>136</td>
<td>137</td>
<td>138</td>
<td>139</td>
<td>140</td>
<td>141</td>
</tr>
<tr>
<td>142</td>
<td>143</td>
<td>144</td>
<td>145</td>
<td>146</td>
<td>147</td>
<td>148</td>
<td>149</td>
<td>150</td>
<td>151</td>
<td>152</td>
<td>153</td>
</tr>
<tr>
<td>154</td>
<td>155</td>
<td>156</td>
<td>157</td>
<td>158</td>
<td>159</td>
<td>160</td>
<td>161</td>
<td>162</td>
<td>163</td>
<td>164</td>
<td>165</td>
</tr>
<tr>
<td>166</td>
<td>167</td>
<td>168</td>
<td>169</td>
<td>170</td>
<td>171</td>
<td>172</td>
<td>173</td>
<td>174</td>
<td>175</td>
<td>176</td>
<td>177</td>
</tr>
<tr>
<td>178</td>
<td>179</td>
<td>180</td>
<td>181</td>
<td>182</td>
<td>183</td>
<td>184</td>
<td>185</td>
<td>186</td>
<td>187</td>
<td>188</td>
<td>189</td>
</tr>
</tbody>
</table>

An atomic mass given in parentheses at the mass number of the isotope of longest half-life for that element.

**Lanthanide series**

<table>
<thead>
<tr>
<th>35</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>45</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td>59</td>
<td>60</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
<td>65</td>
<td>66</td>
<td>67</td>
<td>68</td>
<td>69</td>
<td>70</td>
</tr>
<tr>
<td>71</td>
<td>72</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>76</td>
<td>77</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>81</td>
<td>82</td>
</tr>
<tr>
<td>83</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>92</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>95</td>
<td>96</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>104</td>
<td>105</td>
<td>106</td>
</tr>
<tr>
<td>107</td>
<td>108</td>
<td>109</td>
<td>110</td>
<td>111</td>
<td>112</td>
<td>113</td>
<td>114</td>
<td>115</td>
<td>116</td>
<td>117</td>
<td>118</td>
</tr>
<tr>
<td>119</td>
<td>120</td>
<td>121</td>
<td>122</td>
<td>123</td>
<td>124</td>
<td>125</td>
<td>126</td>
<td>127</td>
<td>128</td>
<td>129</td>
<td>130</td>
</tr>
<tr>
<td>131</td>
<td>132</td>
<td>133</td>
<td>134</td>
<td>135</td>
<td>136</td>
<td>137</td>
<td>138</td>
<td>139</td>
<td>140</td>
<td>141</td>
<td>142</td>
</tr>
<tr>
<td>143</td>
<td>144</td>
<td>145</td>
<td>146</td>
<td>147</td>
<td>148</td>
<td>149</td>
<td>150</td>
<td>151</td>
<td>152</td>
<td>153</td>
<td>154</td>
</tr>
<tr>
<td>155</td>
<td>156</td>
<td>157</td>
<td>158</td>
<td>159</td>
<td>160</td>
<td>161</td>
<td>162</td>
<td>163</td>
<td>164</td>
<td>165</td>
<td>166</td>
</tr>
<tr>
<td>167</td>
<td>168</td>
<td>169</td>
<td>170</td>
<td>171</td>
<td>172</td>
<td>173</td>
<td>174</td>
<td>175</td>
<td>176</td>
<td>177</td>
<td>178</td>
</tr>
<tr>
<td>179</td>
<td>180</td>
<td>181</td>
<td>182</td>
<td>183</td>
<td>184</td>
<td>185</td>
<td>186</td>
<td>187</td>
<td>188</td>
<td>189</td>
<td>190</td>
</tr>
</tbody>
</table>

**Actinide series**

<table>
<thead>
<tr>
<th>35</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>45</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td>59</td>
<td>60</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
<td>65</td>
<td>66</td>
<td>67</td>
<td>68</td>
<td>69</td>
<td>70</td>
</tr>
<tr>
<td>71</td>
<td>72</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>76</td>
<td>77</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>81</td>
<td>82</td>
</tr>
<tr>
<td>83</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>92</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>95</td>
<td>96</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>104</td>
<td>105</td>
<td>106</td>
</tr>
<tr>
<td>107</td>
<td>108</td>
<td>109</td>
<td>110</td>
<td>111</td>
<td>112</td>
<td>113</td>
<td>114</td>
<td>115</td>
<td>116</td>
<td>117</td>
<td>118</td>
</tr>
<tr>
<td>119</td>
<td>120</td>
<td>121</td>
<td>122</td>
<td>123</td>
<td>124</td>
<td>125</td>
<td>126</td>
<td>127</td>
<td>128</td>
<td>129</td>
<td>130</td>
</tr>
<tr>
<td>131</td>
<td>132</td>
<td>133</td>
<td>134</td>
<td>135</td>
<td>136</td>
<td>137</td>
<td>138</td>
<td>139</td>
<td>140</td>
<td>141</td>
<td>142</td>
</tr>
<tr>
<td>143</td>
<td>144</td>
<td>145</td>
<td>146</td>
<td>147</td>
<td>148</td>
<td>149</td>
<td>150</td>
<td>151</td>
<td>152</td>
<td>153</td>
<td>154</td>
</tr>
<tr>
<td>155</td>
<td>156</td>
<td>157</td>
<td>158</td>
<td>159</td>
<td>160</td>
<td>161</td>
<td>162</td>
<td>163</td>
<td>164</td>
<td>165</td>
<td>166</td>
</tr>
<tr>
<td>167</td>
<td>168</td>
<td>169</td>
<td>170</td>
<td>171</td>
<td>172</td>
<td>173</td>
<td>174</td>
<td>175</td>
<td>176</td>
<td>177</td>
<td>178</td>
</tr>
<tr>
<td>179</td>
<td>180</td>
<td>181</td>
<td>182</td>
<td>183</td>
<td>184</td>
<td>185</td>
<td>186</td>
<td>187</td>
<td>188</td>
<td>189</td>
<td>190</td>
</tr>
</tbody>
</table>

An atomic mass given in parentheses at the mass number of the isotope of longest half-life for that element.
### Names, Formulas, and Charges of Some Common Ions

#### Positive Ions (Cations)
- aluminum: $\text{Al}^{3+}$
- ammonium: $\text{NH}_4^+$
- barium: $\text{Ba}^{2+}$
- cadmium: $\text{Cd}^{2+}$
- calcium: $\text{Ca}^{2+}$
- chromium(II): $\text{Cr}^{2+}$
- chromium(III): $\text{Cr}^{3+}$
- cobalt: $\text{Co}^{2+}$
- copper(I): $\text{Cu}^{+}$
- copper(II): $\text{Cu}^{2+}$
- hydrogen, hydronium: $\text{H}^+$, $\text{H}_3\text{O}^+$
- iron(II): $\text{Fe}^{2+}$
- iron(III): $\text{Fe}^{3+}$
- lead(II): $\text{Pb}^{2+}$
- lithium: $\text{Li}^+$
- magnesium: $\text{Mg}^{2+}$
- manganese(II): $\text{Mn}^{2+}$
- mercury(I): $\text{Hg}^{2+}$
- mercury(II): $\text{Hg}^{2+}$
- nickel: $\text{Ni}^{2+}$
- potassium: $\text{K}^+$
- scandium: $\text{Sc}^{3+}$
- silver: $\text{Ag}^+$
- sodium: $\text{Na}^+$
- strontium: $\text{Sr}^{2+}$
- tin(II): $\text{Sn}^{2+}$
- tin(IV): $\text{Sn}^{4+}$
- zinc: $\text{Zn}^{2+}$

#### Negative Ions (Anions)
- acetate: $\text{CH}_3\text{COO}^-$
- bromide: $\text{Br}^-$
- carbonate: $\text{CO}_3^{2-}$
- hydrogen carbonate, bicarbonate: $\text{HCO}_3^-$
- chloride: $\text{Cl}^-$
- chlorate: $\text{ClO}_3^-$
- chromate: $\text{CrO}_4^{2-}$
- dichromate: $\text{Cr}_2\text{O}_7^{2-}$
- fluoride: $\text{F}^-$
- hydride: $\text{H}^-$
- hydroxide: $\text{OH}^-$
- hypochlorite: $\text{ClO}^-$
- iodide: $\text{I}^-$
- nitrite: $\text{NO}_2^-$
- oxalate: $\text{C}_2\text{O}_4^{2-}$
- oxide: $\text{O}_2^-$
- perchlorate: $\text{ClO}_4^-$
- permanganate: $\text{MnO}_4^-$
- phosphate: $\text{PO}_4^{3-}$
- monohydrogen phosphate: $\text{HPO}_4^{2-}$
- dihydrogen phosphate: $\text{H}_2\text{PO}_4^-$
- sulfate: $\text{SO}_4^{2-}$
- hydrogen sulfate, bisulfate: $\text{HSO}_4^-$
- sulfide: $\text{S}^{2-}$
- hydrogen sulfide, bisulfide: $\text{HS}^-$
- sulfite: $\text{SO}_3^{2-}$
- hydrogen sulfite, bisulfite: $\text{HSO}_3^-$
<table>
<thead>
<tr>
<th>Name</th>
<th>Structural Shape</th>
<th>Atoms Bonded to Central Atom</th>
<th>Lone Pairs of Electrons</th>
<th>Bond Angle</th>
<th>Sample Formula</th>
<th>Electron Dot Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td><img src="image" alt="Linear Structure" /></td>
<td>2</td>
<td>0</td>
<td>180°</td>
<td>BeH₂</td>
<td>H:Be:H</td>
</tr>
<tr>
<td>trigonal planar</td>
<td><img src="image" alt="Trigonal Planar Structure" /></td>
<td>3</td>
<td>0</td>
<td>120°</td>
<td>BF₃</td>
<td><img src="image" alt="BF₃ Structure" /></td>
</tr>
<tr>
<td>Tetrahedral</td>
<td><img src="image" alt="Tetrahedral Structure" /></td>
<td>4</td>
<td>0</td>
<td>109.5°</td>
<td>CH₄</td>
<td><img src="image" alt="CH₄ Structure" /></td>
</tr>
<tr>
<td>trigonal pyramidal</td>
<td><img src="image" alt="Trigonal Pyramidal Structure" /></td>
<td>3</td>
<td>1</td>
<td>107°</td>
<td>NH₃</td>
<td><img src="image" alt="NH₃ Structure" /></td>
</tr>
<tr>
<td>Bent</td>
<td><img src="image" alt="Bent Structure" /></td>
<td>2</td>
<td>2</td>
<td>104.5°</td>
<td>H₂O</td>
<td><img src="image" alt="H₂O Structure" /></td>
</tr>
</tbody>
</table>
Appendix 5 – Desperately Seeking Solution worksheet
Desperately Seeking Solution!

Name:__________ Period:_____ Date:_____

Flame Test: Color ________ Possible Negative Ions: OH⁻ CO₃²⁻ SO₄²⁻ NO₃⁻ Cl⁻ PO₄³⁻

Positive Ion ________

USE ONLY THE TESTS YOU NEED TO FIGURE OUT THE NEGATIVE IONS. SHOW WORK!

Physical Properties: Reaction with acids: Reactions with bases:

Reaction / elements: Reaction with salts: Reactions / Compounds: Reactions / Bases:

Net Ionic Equations Work Space:

Solve _______ ❄️ Number of attempts __ Score _______

Solubility of ionic compounds in water

<table>
<thead>
<tr>
<th>Name of ion(x)</th>
<th>Generally Soluble (NVR)</th>
<th>Insoluble (PPT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sodium, potassium, ammonium, lithium</td>
<td>All are soluble</td>
<td></td>
</tr>
<tr>
<td>Chloride, Bromide, Iodide</td>
<td>Rest are soluble</td>
<td>Ag⁺ Hg⁺ Pb⁺²</td>
</tr>
<tr>
<td>Sulfate</td>
<td>Rest are soluble</td>
<td>Ba⁺² Hg⁺ Pb⁺²</td>
</tr>
<tr>
<td>Nitrate</td>
<td>All are soluble</td>
<td></td>
</tr>
<tr>
<td>Acetate</td>
<td>All are soluble</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of compound</th>
<th>Soluble (NVR)</th>
<th>Generally Insoluble (PPT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sulfides</td>
<td>NH₄⁺ Na⁺ K⁺ Mg²⁺ Ca²⁺</td>
<td>Rest are insoluble</td>
</tr>
<tr>
<td>hydroxides</td>
<td>Li⁺ K⁺ Na⁺ Ba⁺²</td>
<td>Rest are insoluble</td>
</tr>
<tr>
<td>carbonates, phosphates</td>
<td>NH₄⁺ Na⁺ K⁺</td>
<td>Rest are insoluble</td>
</tr>
</tbody>
</table>
Appendix 6 – Second semester pencil and paper final exam
2ndfinalDOC  WRITE ON THIS TEST:  SPRING 2000 FINAL
NAME: ____________________  DATE: __________  PERIOD: ________
CALCULATE AND CONVERT TO SCIENTIFIC NOTATION:
1. (6.50 X 10^45) (6.00 X 10^-34)  2. 6.51 X 10^23
   (6 X 10^-9 ) (5 X 10^-13)

SOLVE and ADJUST ANSWERS USING THE PROPER NUMBER OF SIGNIFICANT FIGURES. (convert to scientific notation):
3. 5.00014 X 0.0080  4. 62.0 X 360.0

WRITE THE FORMULAS FOR THE FOLLOWING MOLECULES. ADJUST FOR PROPER VALENCE
5. barium hydroxide  6. ammonium nitrate  7. Carbon tetrafluoride
WRITE THE NAMES FOR THE FOLLOWING FORMULAS. BE SURE TO USE PROPER RULES
8. LiSO4  9. Mg3(PO4)2  10. PBrs

BALANCE THE FOLLOWING EQUATIONS. ID. #11. 12 (IV) Endothermic (X) Exothermic
11. H2O (l) + CO2 (g) -> CO3^2- (aq) + H^+ (aq)  2.2 kJ = Δ H°
12. CaC2 (s) + O2 (g) -> H2O (l) + CO2 (g)  2219.9 kJ = Δ H°
13. SiO2 (s) + C (s) -> CO (g) + Si (s)
14. P4 (s) + 3Cl2 (g) -> 2PCl3 (l)

CHOOSE THE LETTER OF WHICH FACTOR OR BRIDGE WILL CONVERT THE FOLLOWING.
15. 2.9 liters of Ss --→ # mol of S  M. molar mass (g) N. 1 mole  O. # mol Cu  P. 22.4 liters
16. 1.7 mol. KCl --→ #molecules KCl  # mol. Fe  1 mole
17. 1.5 mol. MgSO4 --→ #Liters. MgSO4  Q. 6.02 X 10^23  R. 1 mole  S. #mol Fe  T. 1 mole
18. 2 mol Cu --→ #mol Fe  6.02 X 10^23  #mol Cu  22.4 liters

USE ARROWS (→ OR ← OR "no change") TO INDICATE THE EFFECT OF THE FOLLOWING ON EQUILIBRIUM.
19-21
19. decrease in concentration of H2S  20. decrease in temperature  21. increase in temperature

USE ARROWS (↑ OR ↓) TO INDICATE THE EFFECT OF THE FOLLOWING ON REACTION RATE.
22-24
22. increase in concentration of H2  23. decrease in temperature  24. increase in temperature

Calculate pH problems. show all work, formulas, and attach answer
25. Calculate pH of a strong acid at 1.75 X 10^-4 M

26. Calculate pOH of a strong base at 1.75 X 10^-2 M

EQUATIONS:  Greek Prefixes:
pH = - log [H3O^+]  mono one
pOH = 14-(- log [H3O^+])  di two
[H3O^+] = antilog - pH  tri three
tetra four
penta five
hexa six
Calculate Oxidation Numbers for the following:

27. $SO_3^2-$  
$S = 4$  
$O = \frac{(-2) \cdot 3}{2} = -6$  
$I = -2$

28. $Br_2$  
$Br = 0$

29. $CF_4$  
$C = 4$  
$F = -1$  
$I = (-1) = -4$

DRAW ELECTRON DOT STRUCTURES FOR THE FOLLOWING COMPOUNDS OR POLYATOMIC IONS

30. $BrO_4^-$  

31. $Na^+$  

32. $F_2$

33. What group of elements do the lowest points of each period represent?
   a) halogens  b) noble gases  c) group 1A  d) alkaline earth metals. What group

34. What group of elements do the highest points of each period represent?
   a) halogens  b) noble gases  c) group 1A  d) alkaline earth metals

35. What "generally" happens to the amount of ionization energy for each element as you move across a period?
   a) increases  b) decreases  c) stays the same

36. What "generally" happens to the amount of ionization energy for each element as you move down a group?
   a) increases  b) decreases  c) stays the same

37. What group of elements do second and third lowest points of each period represent?
   a) halogens  b) noble gases  c) group 1A  d) alkaline earth metals. What group

38. What group of elements do the highest points of each period represent?
   a) halogens  b) noble gases  c) group 1A  d) alkaline earth metals

39. What "generally" happens to the size or atomic radius for each element as you move across a period?
   a) increases  b) decreases  c) stays the same

40. What "generally" happens to the size or atomic radius for each element as you move down a group?
   a) increases  b) decreases  c) stays the same

41. Cut out of test
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>43. Bond with unequal sharing of electrons</td>
<td>A. Polar covalent bond</td>
</tr>
<tr>
<td>44. Measures average kinetic motion of particles</td>
<td>B. Quantum mechanics</td>
</tr>
<tr>
<td>45. Amount of energy required to remove an electron from an atom</td>
<td>C. Ionization energy</td>
</tr>
<tr>
<td>46. Atoms of the same element with different numbers of neutrons</td>
<td>D. Isotopes</td>
</tr>
<tr>
<td>47. The number of protons or electrons</td>
<td>E. Crystal lattice</td>
</tr>
<tr>
<td>48. Substance that gives up hydrogen protons in $\text{H}_2\text{O}$</td>
<td>F. Nonmetals</td>
</tr>
<tr>
<td>49. Examples include neon, helium and xenon</td>
<td>G. Noble gases</td>
</tr>
<tr>
<td>50. Substance that accepts hydrogen protons in $\text{H}_2\text{O}$</td>
<td>H. Concentration</td>
</tr>
<tr>
<td>51. Examples include bromine, chlorine, oxygen</td>
<td>I. Temperature</td>
</tr>
<tr>
<td>52. This is a measure of number of moles per liter</td>
<td>J. Base</td>
</tr>
<tr>
<td>53. Determining concentrations of substances using chemical reactions with other substances</td>
<td>K. Acid</td>
</tr>
<tr>
<td>54. Packet of energy released by excited electron</td>
<td>L. Atomic number</td>
</tr>
<tr>
<td>55. 3-D grouping of anions and cations</td>
<td>M. Ionic bond</td>
</tr>
<tr>
<td>56. Bond formed by electrostatic attraction</td>
<td>N. Titration</td>
</tr>
<tr>
<td>57. Determines physical arrangement of atomic structures</td>
<td>O. Photon</td>
</tr>
<tr>
<td>58. Ratio of equilibrium concentrations of products over reactants</td>
<td>P. Mass number</td>
</tr>
<tr>
<td>59. Quality of being dissolved or dissolving</td>
<td>Q. $K_{\text{eq}}$</td>
</tr>
<tr>
<td>60. Measures affinity for electrons</td>
<td>R. Equilibrium</td>
</tr>
<tr>
<td>61. Separation of ionic compounds into ions by hydration</td>
<td>S. Saturated</td>
</tr>
<tr>
<td>62. Reaction where a substance loses electrons</td>
<td>T. Alkali metals</td>
</tr>
<tr>
<td>63. Reaction where a substance gains electrons</td>
<td>U. Le Chatelier's principle</td>
</tr>
<tr>
<td>64. A new solid formed in a double replacement reaction</td>
<td>V. Solubility</td>
</tr>
<tr>
<td>65. Chemical that shows a color change depending on hydronium ion concentration</td>
<td>W. Ground state</td>
</tr>
<tr>
<td>66. System where the rate of both directions of the reaction are equal</td>
<td>X. Indicator</td>
</tr>
<tr>
<td>67. Conducts electricity when hydrated</td>
<td>Y. Dissociation</td>
</tr>
<tr>
<td>68. Soft metals that react violently with water</td>
<td>Z. Electronegativity</td>
</tr>
<tr>
<td>69. Energy level with the least amount of energy</td>
<td>A. Electrolyte</td>
</tr>
<tr>
<td>70. Solution that can't dissolve more solute</td>
<td>B. Precipitate</td>
</tr>
<tr>
<td>71. States that when a factor disturbs an equilibrium system, the system responds to partially undo what has been done</td>
<td>C. Reduction</td>
</tr>
<tr>
<td>72. Total of protons and neutrons</td>
<td>D. Oxidation</td>
</tr>
</tbody>
</table>
Appendix 7 – Final Student Survey
IMMEX Labs / The Massachusetts Institute of Technology

Student Survey 2

As before, this is a confidential questionnaire. Your individual answers will be held in confidence by the researcher. Please answer each question as honestly as you can. Thanks!

Student IMMEX ID: ____________________________

1. What is your birth date? (mm/dd/yy) _____/_____/

When solving IMMEX problems how often did you

2. Use knowledge you learned during the school year? Not Very Some Quite A lot
   at all little Some a bit A lot
   1 2 3 4 5

3. Guess at an answer before collecting data? 1 2 3 4 5

4. Guess the correct answer without knowing how to actually solve the problem? 1 2 3 4 5

How effective has IMMEX been in

5. Helping you learn Chemistry facts (for example, how to write net ionic equations)? 1 2 3 4 5

6. Helping you apply Chemistry facts (for example, how to actually use solubility rules for something)? 1 2 3 4 5

7. Helping you understand Chemistry (for example, to predict the products of mixing two compounds together before actually doing so)? 1 2 3 4 5

8. Evaluating what you really know about Qualitative Chemistry (net ionic equations, product prediction, Flame tests, solubility, pH, etc.)? 1 2 3 4 5

9. How much did you like using IMMEX? 1 2 3 4 5

10. How much do you think using IMMEX improves your problem solving skills? 1 2 3 4 5

11. How often do you play video games? 1 2 3 4 5
12. If you had to grade yourself on how well you know Qualitative Chemistry (net ionic equations, product prediction, Flame tests, solubility, pH, etc.), what grade would you assign yourself?

In general, how often do you do the following when solving problems:

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Look at the available information before making a final decision?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1.</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>14. Pick the first solution that comes to mind?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>15. Pick the solution that seems easiest?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>16. Try to look for more than one solution to the problem?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>17. Look for information that will eliminate a solution?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>18. Look for information that will support a solution?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>19. Ignore information that goes against what you think the solution is?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>20. Only use information that goes along with what you think the solution is?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>21. Talk to your teacher or other people to get information?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>22. Routinely ask yourself what information is important?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Very Often</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Thanks for taking the time to answer this survey!

- Have a great Summer-
Appendix 8 – The conversion of IRT theta to odds of problem solving

The Rasch “single parameter” model used in Item Response theory is

\[ \Pr(\text{Solve}) = \frac{1}{1 + e^{-\theta}} \]

where: \(\theta\) is a measure of student ability,
\(e\) is the natural number,
and \(\Pr(\text{Solve})\) is the probability a student with ability \(\theta\) will solve the problem.

Therefore:

\[
1 + e^{-\theta} = \frac{1}{\Pr(\text{Solve})}
\]

\[
e^{-\theta} = \frac{1}{\Pr(\text{Solve})} - 1 = \frac{1 - \Pr(\text{Solve})}{\Pr(\text{Solve})}
\]

\[e^\theta = \frac{\Pr(\text{Solve})}{1 - \Pr(\text{Solve})} = \text{odds (Solve)}\]

\[\theta = \ln(\text{odds (Solve)})\]
Appendix 9 – The Hazmat IMMEX™ Problem Space

The "Big One" has arrived...
An earthquake has just shaken up a local school causing many chemicals to fall off of the shelves in the chemical stockroom, spilling a number of compounds onto the floor. Many of these chemicals can be harmful or dangerous if not handled and disposed of properly, and some combinations of chemicals can produce hazardous fumes or violent reactions. The water pipe supplying the chemistry classrooms has also broken, and water is beginning to flow into the storeroom.

Your task as the HazMat team chemical analyst is to identify the chemicals which have spilled in order to arrange for safe clean-up. It has been estimated that the water will reach the spilled chemicals in twenty-four hours, so you have only ten hours (600 minutes) to identify the spilled chemicals before the danger of reactions becomes serious.
Bibliography


Drake University (2000). Notes on validity and reliability. www.educ.drake.edu, Drake University School of Education.


IAEA (1994). Third International math and science Study (TIMMS), International Association for the evaluation of Educational Achievement.


Sanders, W. L. and S. P. Horn *Educational Assessment Reassessed: The Usefulness of Standardized and Alternative Measures of Student Achievement as Indicators for the Assessment of Educational Outcomes*. http://olam.ed.asu.edu/epaa/v3n6.html, Arisona State University, Tempe.


