2000

Performance Assessment and English Skill.

Zhenghe Zhang
Louisiana State University and Agricultural & Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_disstheses

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_disstheses/7242

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Historical Dissertations and Theses by an authorized administrator of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
INFORMATION TO USERS

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

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

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

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

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

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
PERFORMANCE ASSESSMENT AND ENGLISH SKILL

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Educational Leadership, Research and Counseling

by

Zhenghe Zhang
B.A., Shanghai Normal University, 1983
M.A., Louisiana State University, 1990
May 2000
ACKNOWLEDGMENTS

I am extremely grateful to my major professor Dr. Eugene Kennedy for his valuable support, guidance and patience throughout my work at LSU. I would like to thank my committee members Dr. A. Tashakkori, Dr. W. Davis, Dr. R. Fossey. Taylor, and Dr. J. Kotrlik for their invaluable suggestions, comments and corrections.

I would also like to thank my wife and our son for cheering me throughout.
TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................ ii

LIST OF TABLES ............................................................................................................. v

LIST OF FIGURES .......................................................................................................... vi

ABSTRACT ...................................................................................................................... vii

CHAPTER I INTRODUCTION ......................................................................................... 1
  Introduction ................................................................................................................ 1
  Background ................................................................................................................. 4
  Writing Assessment ................................................................................................. 13
  Writing Assessment and English as Second Language Students ....................... 17
  Statement of Problem and Research Questions ..................................................... 19
  Significance ............................................................................................................. 19

CHAPTER II REVIEW OF RELATED LITERATURE ....................................................... 21
  Definitions of Assessment Approaches ................................................................. 21
  Sources for Performance Assessment .................................................................. 23
    The Necessity for a Developmental Perspective ............................................... 23
    The Emergence of a Symbol-System Perspective ............................................. 24
    The Decomposability and the Decontextualization of Assumptions .................. 26
  Evidence for the Existence of Multiple Features ................................................... 29
  Recognition of Individual Differences ................................................................. 31
  The Desirability of Assessing Learning in Context .................................................. 31
  General Features of Performance-Based Assessment ........................................... 32
  Criteria of Comparisons between Multiple-Choice and Performance-Based Formats ......................................................................................................................... 36
    Success Rate ....................................................................................................... 36
    The Traits Measured by the Test ....................................................................... 38
    Retention ............................................................................................................ 41
    Strategy Choice ................................................................................................. 43
    Diagnostic Information ....................................................................................... 45
    Writing Assessment ........................................................................................... 50
  Summary .................................................................................................................. 55

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>CHAPTER III METHODOLOGY</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>60</td>
</tr>
<tr>
<td>Instruments and Data</td>
<td>60</td>
</tr>
<tr>
<td>Sample</td>
<td>62</td>
</tr>
<tr>
<td>Definition of Regular Student</td>
<td>63</td>
</tr>
<tr>
<td>Definition of Limited English Proficient (LEP) Student</td>
<td>63</td>
</tr>
<tr>
<td>Analyses</td>
<td>63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER IV RESULTS</th>
<th>69</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results for Hypothesis 1</td>
<td>69</td>
</tr>
<tr>
<td>Results for Hypothesis 2</td>
<td>72</td>
</tr>
<tr>
<td>Results for Hypothesis 3</td>
<td>74</td>
</tr>
<tr>
<td>Results for Hypothesis 4</td>
<td>77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER V DISCUSSION AND CONCLUSIONS</th>
<th>89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion</td>
<td>89</td>
</tr>
<tr>
<td>Conclusions</td>
<td>92</td>
</tr>
<tr>
<td>Implications</td>
<td>93</td>
</tr>
</tbody>
</table>

| REFERENCES | 97 |

| VITA | 110 |
LIST OF TABLES

1. Mean Dimension Scores for English the Written Composition Test .......... 70
2. Analysis of Variances for Student Group for the Written Composition Test ..................................................................... 72
3. Mean Subskill Scores for the English Language Arts Test .................. 73
4. Analysis of Variances for Student Group for the English Language Arts Test ................................................... 75
5. Mean Differences across PBA and MC .............................................................. 76
6. Pearson Correlations among 13 Subtests for Regular Students .............. 78
7. Oblique Two-Factor Solutions for the 13 subtest –data for Regular Students .................................................................................... 79
8. Pearson Correlations among 13 Subtests for LEP Students ................. 81
9. Oblique Two-Factor Solution for the 13 subtest-data for LEP Students .......................................................................................... 82
10. Summary of Confirmatory Factor Analysis .............................................. 83
11. Simultaneous Tests for the Invariance of Test Format Structure .......... 87
LIST OF FIGURES

1. Baseline Model of Test Format for Regular Education Students ............... 85
2. Baseline Model of Test of Format for LEP Students ............................. 86
ABSTRACT

Recent developments in the philosophy of validity, highlighting the importance of investigating the consequences of assessment use, provide theoretical support for the move toward performance assessment. Numerous issues related to the choice of assessment approach (e.g., multiple-choice or performance-based) are not well understood today. This study investigated the potential interaction between assessment approach and the English skill of the examinee and the similarity of the information provided by multiple-choice and essay writing tests.

Four hypotheses were examined in this study. Hypothesis 1 results indicated that students who were native speakers of English (regular students) achieved higher scores on the writing test than did the LEP students. Regular students were significantly different from LEP students in support/elaboration/organizations, sentence structure, and usage. However, there were no significant differences between these groups in writing mechanics and responsiveness to assignment. The results for Hypothesis 2 indicated that regular students scored significantly higher in all the eight subskills of the multiple-choice test than did the LEP students. Hypothesis 3 indicated a substantial interaction between English skill and assessment approach. The differences between the two groups on the higher-order elements of composition were much greater on the writing test than the multiple-choice test. Finally, the results for hypothesis 4 indicated that the factorial structure of the two tests were essentially identical for the two groups.
It was concluded that the multiple-choice test consistently differentiated between the two groups of students studied in the direction consistent with the English skill levels of the students. In contrast, the performance-based writing test yielded expected differences for higher-order skills, but failed to separate the groups on lower level writing skills. This suggests that the multiple-choice test is more general in its utility than is true of the performance-based writing test. Also, it suggests that the writing test may be superior to the multiple-choice test at differentiating students on the higher-order elements of compositions. Conversely, it may be the case that the writing test exaggerates the differences between the groups raising the possibility of bias. The implications of this for high-stakes testing programs are discussed and recommendations are offered.
CHAPTER 1

INTRODUCTION

Introduction

Objective testing, typified by the multiple-choice item format, has been the predominant approach to educational measurement and assessment for several decades. Recently, however, there has been a significant trend toward the use of performance-based assessments as supplements or even replacements for traditional objective tests (Linn, 1991). Examples of performance-based assessments include open-ended problems in mathematics, language arts, and social studies; essays which mimic communication in real-world settings; hands-on science experiments; computer simulations; and portfolios of student work. Collectively, such measures are frequently referred to as "authentic" assessments because they involve the performance of tasks that are valued in their own right (Archibald & Newman, 1988; Wiggins, 1989). In contrast, paper-and-pencil multiple-choice tests derive their value primarily as indicators or correlates of valued performances.

Indicators are often confused with goals, just as norms are often confused with standards. When this happens, there is a possibility that the indicator may supplant the goal as the focus of attention. In the context of testing, this distortion occurs when performance on a traditional objective test drives the educational agenda. Students may be taught to answer items on 'the test' correctly, yet they may have little exposure to, or mastery of, the broader content of which the objective test is only a sample. In recent years, the potential for this distortion has increased as the number of mandatory high

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
stakes assessments in education has increased. This trend has helped motivate pleas for performance-based or alternative assessments (Linn, 1991).

Although the recent push for performance-based assessments seems new to many, in fact, this has been standard advice from measurement specialists for several decades. Lindquist (1951), for example, argued that "... it should always be the fundamental goal of the achievement test constructor to make the elements of his test series as nearly equivalent to, or as much like, the elements of the criterion series as consequences of efficiency, comparability, economy, and expediency will permit" (p. 152). With regard to the construction of tests intended to measure higher order thinking and critical reasoning skills, Lindquist (1951) went on to note: "The most important consideration is that the test questions require the examinee to do the same things, however complex, that he is required to do in the criterion situations (p. 154)".

Lindquist's views on the proper goals for the constructor of achievement tests add historical perspective to the current debate on alternative assessment instruments. More contemporary writers, echoing his primary position, have described significant instructional limitations of multiple-choice tests (e.g., Archibald & Newman, 1988; Frederiksen, 1984; Resnick & Resnick, 1991; Shepard, 1991). Others point to advances in cognitive and developmental psychology which suggest a discordance between the multiple-choice format and the ways in which students think and learn (e.g., Glaser, 1984, Greeno, 1989; Snow 1980; Snow & Lohman, 1989). Still other researchers and policymakers have called for reform and accountability, focusing attention on valued performances (i.e., what children know, don't know, and should know) rather than mere indicators (Office of Educational Research and Improvement, 1990).
Performance assessments are increasingly being used in large scale high-stakes testing programs. Examples of nationwide testing and assessment programs that include performance items are the College Board Advanced Placement tests, the ACT College Outcome Measures Programs, and Workkeys Assessment (Zwick, Donoghue, & Grima, 1993). Many states in the United States are administering or planning complete performance assessments in areas such as writing, reading, mathematics, and social studies (Linn, 1991).

With widespread and influential support for alternatives to traditional objective tests, it is not premature to pose questions about the validity criteria that alternative assessments ought to satisfy (Linn, 1993). Is it sufficient, for example, to establish that a performance task appears to mimic its real-world counterpart? Or should researchers empirically establish that the constructs measured by these tasks are similar across cultural contexts and ability levels; that they provide representative samples of content domains; or, that they are free of cultural biases? These and related issues have been the focus of much attention in recent years and they constitute the chief concern of this study. In particular, this study is concerned with the possibility of an interaction between assessment format and the skill level of examinees in the context of writing assessment.

This chapter presents the background to this study. The first part traces the historical predominance of performance-based and objective assessments in this country. The next section presents a discussion of the issues of validity and reliability as they have come to be the focus of much attention. The final sections focus on writing assessment and the specific research questions that guided this effort.
Background

In the early universities, and up to the nineteenth century, examinations were usually oral (Scates, 1947). In this country, when enrollments were small, the final examinations of eighth graders were used as a community social event in which everyone turned out to hear the "scholars" put through their paces. As compulsory attendance laws were introduced, this practice of oral examinations became impossible (Downie, 1958). In the 1840s Horace Mann, who was then secretary of the Massachusetts Board of Education, wrote enthusiastically of written examinations that were in use in Boston. Because these examinations appeared to be efficient and to bring needed conformity to student assessments, the use of written examinations soon spread.

Not many years after their introduction, written essays were the dominant type of assessment used to evaluate student learning. However, beginning in this century several researchers started to question their quality (see Downie, 1958). Starch and Elliott (1912), for example, conducted a pioneering study of English composition essays. They had two English composition papers, determined to be of equal merit, graded by 142 English teachers. Grades on the first paper ranged between 50 and 98 with a median of 80.2 whereas grading on the second paper ranged between 64 and 99 with a median of 88.2. Starch and Elliot attributed this ambiguous outcome to the 'subjective' nature of English assessment. In a later study, these same authors had 138 teachers grade a plane geometry paper (Starch and Elliot, 1913). In this instance, grades ranged between 28 and 95 with a normal spread throughout the range.

Results similar to those of Starch and Elliot were replicated in several other studies. One such study was conducted by Falls (1928) who asked 100 English teachers
to grade a paper written by a high-school senior and to estimate the grade level of the writer of the essay. In fact, the essay was selected by a committee as being a 'first-class' paper, and the writer of the essay was a correspondent for a large metropolitan daily. Grades on this paper ranged from 60 to 98, and the writer of the paper was said to be at levels ranging from fifth grade to a junior in college.

The work of Starch and Elliot and others raised questions about the essay as a technique for evaluating student achievement. More specifically, these studies questioned the quality of the grading of these examinations—i.e., a problem of reliability. Other challenges to the essay, as an assessment tool, have focused on the adequacy with which a content area is sampled, the efficiency with which skills are measured, and the potential for bias in evaluations (Downie, 1958).

While alternatives to essays and oral examinations have always existed, it was not until the work of Rice (1987) that the modern movement toward 'objective' testing started to achieve momentum. An educational researcher, Rice (1897) was concerned with achievement in spelling as it related to the amount of time given to spelling instruction (Scates, 1963). Having constructed spelling tests and administered them in various large cities, Rice found that the amount of time spent in teaching spelling had no effect upon achievement. The significance of this work was that it was the first time that a researcher successfully studied a large number of individuals with a single test (Downie, 1958).

Following Rice's work, large scale 'standardized' assessments were used by the US military during World War I in an effort to more effectively make use of the talents of recruits. Not long after this high profile effort, large scale standardized assessment moved
into the educational and business arenas. Driven by the focus on efficiency and objectivity, the multiple choice item was the format of choice, all but rendering the 'inefficient' and 'subjective' essay obsolete (Haney and Madaus, 1989).

With the development of fast and highly efficient scoring machines (Linquist, 1954), testing with multiple-choice items was possible on a scale that would be impossible with tests requiring written responses. In the 1981-1982 academic year, for example, nearly a million and a half students took the Scholastic Aptitude Test (SAT), and more than a quarter of a million took one or more of the College Board achievement tests. The American College Testing Program tested nearly a million candidates in the same period (Frederiksen, 1984).

Since the total number of students enrolled in schools (K through 12) in the United States is in the neighborhood of 46 million, the impact of testing is substantial. Moreover, teachers are routinely trained to develop and use multiple-choice items in their classrooms, and many colleges have testing bureaus to help faculty cast their tests in the multiple-choice form. The economical machine-scored tests had, for all practical purposes, driven out other types of examinations, at least for large-scale testing programs (Haney & Madaus, 1989).

As the number of students taking standardized multiple-choice tests has increased and the consequences have become more significant, there has been a corresponding increase in both the volume and frequency of criticisms of the multiple-choice item (Hoffman, 1962; Gross, 1962; Houts, 1971; Gould, 1981; Owen, 1985; Crouse & Trusheim, 1988; & Cannell, 1987). In addition to published complaints, there have been hearings in which standardized testing has been subjected to criticism. Further, since the
1960s there have been hundreds (if not thousands) of lawsuits brought in both state and federal courts concerning testing in education and in employment. Indeed, the complaints and criticisms have been so substantial that some have written of a "war on testing" (Lerner, 1979; Goliath & Gallup, 1980; & Hargaden, 1980).

Of the myriad challenges to the multiple-choice item, one of the most common is that they can only be used to test surface or superficial learning (Green, 1981). Investigating this possibility, Bowman and Peng (1972) asked five psychologists to judge which one of four cognitive abilities was predominantly involved in answering each item in four forms of the GRE Advanced Psychology Test. The four abilities were named memory, comprehension, analytic thinking, and evaluation. Memory was defined as "the simple reproduction of facts, formulae, or other items of remembered content." The consensus of the judges was that 70% of the items were memory items, 15% measured comprehension, 12% required analytic thinking, and only 3% involved evaluation. They concluded that even a carefully developed test designed to measure potential or capacity for graduate study was primarily a measure of factual knowledge (Bowman & Peng, 1972).

A similar study (Levine, McGuire, & Nattress, 1970) was based on a test called the Orthopedic In-Training Examination, a test used to measure competence in orthopedic medicine. Trained judges were asked to sort the items into categories representing the cognitive processes they thought were used by candidates in answering items. More than half of the items were unanimously believed to require only recall of information, while less than 25% were believed by even one judge to require interpretation of data, application of a principle, or evaluation. An effort was then made
to improve the test by increasing the proportion of items that measure higher order skills. Explicit instructions were given to the item writers, and the items were systematically reviewed. It was found after the new test was completed that a majority of the items were still judged to require only recall.

A second area in which multiple-choice items have been criticized is their influence on learning. Gay (1980) compared multiple-choice and short-answer tests with respect to their influence on retention. A "surprise" final examination consisting of multiple-choice and short-answer items was developed. This examination covered the same content as that of the formative tests used by the teacher during instruction. For the multiple-choice part of the examination, there was no difference between student performance on the final and the teaching tests; but for the short-answer part, performance was better for students whose teaching-test format was also free-response. Gay suggests that "one strategy for maximum retention would seem to be the use of [short-answer] testing whenever possible" (p.50).

Voss (1974) was interested in the influence of format on nonspecific transfer of learning from prose. On each of three days students were given a prose passage that was to be read four times, with each reading followed by a test. Performance improved after each reading for both multiple-choice and completion formats. However, there was no day-to-day improvement; in other words, there was no evidence of nonspecific transfer. Further, it was noted that for the later trials on a given day, performance tended to be poorer for students given multiple-choice tests relative to those given the completion items. Voss suggested that the exposure to incorrect options interfered with retention of the correct answers.
Another aspect of the criticism of the impact of multiple-choice items on learning concerns the nature of the constructs which are measured. Frederiksen and Ward (1978), for example, made comparisons of several multiple-choice and free-response item formats. One such test, called Formulating Hypotheses, (Frederiksen & Ward, 1978) was intended to simulate a kind of problem frequently faced by scientists. Each problem consisted of a brief description of a research study in a behavioral science, a graph or table showing the results, and a statement of the major finding. The task was to write possible explanations, or hypotheses, that might account for the finding, including competing hypotheses that ought to be considered. A multiple-choice form was later constructed in which a list of hypotheses was presented for each problem; the list was based on a tabulation of the hypotheses written by students who took the test in free-response form. Scores were obtained that reflected the quality, number, and unusualness of the hypotheses written or chosen, not just the number "correct."

Correlations between corresponding scores for the two formats were found to be very low. For example, for scores reflecting mean quality of the ideas, the correlation between formats was .18, and for scores based on number of ideas, the correlation was .19. Frederiksen and Ward interpreted these findings as evidence that the two formats do not measure the same construct.

To discover more specifically what constructs were measured by the two tests, extension loadings of the scores on various cognitive factors were obtained (Ward, Frederiksen, & Carlson, 1980). It was found that a knowledge factor, based on the GRE Advanced Psychology Test and verbal and reasoning factors had quite similar patterns of correlation with scores for both forms, especially for scores representing the quality of
the hypotheses. The most striking difference had to do with the role of ideational fluency (Guilford, 1967, 1982), which the authors interpreted as skill in broadly searching long-term memory for relevant ideas. None of the scores from the multiple-choice form had appreciable correlation with fluency measures. For the free-response form, however, scores reflecting number of ideas, number of unusual ideas, and the number of ideas that are both unusual and of high quality all had substantial correlations with fluency, even though they were less reliable than the comparable scores based on multiple-choice responses. Only the free-response form appeared to require a broad search for relevant ideas and information stored in memory.

Given the seriousness and range of criticisms of the multiple-choice item format, it is not surprising that the search for alternatives has gained momentum in recent years. While these alternatives have taken many forms (e.g., open-ended items, exhibitions, portfolios, etc.) most permit flexibility in student responses and involve fewer items than is true of conventional multiple-choice tests. The flexibility in responses almost always necessitates scoring by readers or graders and the relatively smaller number of discrete tasks raises questions concerning the adequacy with which a content domain is sampled. Both problems are reminiscent of the concerns raised at the beginning of this century when the essay, a form of open-ended assessment, once before dominated the educational assessment landscape.

Perhaps in an effort to avoid the mistakes of the past, contemporary proponents of performance-based assessments have devoted considerable energy to issues of validity and reliability. Linn (1991) argues that serious validation of alternative assessments needs to include evidence regarding the intended and unintended consequences, the degree to
which performance on specific assessment tasks transfer, and the fairness of the assessments. Further, he posits that evidence is also needed regarding the cognitive complexity of the processes students imply in solving assessment problems and the meaningfulness or fairness of the problems for students and teachers.

Questions of fairness arise not only in the selection of performance tasks but also in the scoring of responses (Haertel, 1987). As Stiggins (1987) has stated, it is critical that scoring procedures are designed to assure that "performance ratings reflect the examinee's true capabilities and are not a function of the perceptions and biases of the persons evaluating the performance" (p.33).

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaranam, 1972; Brennan, 1983; Shavelson, Webb, & Rowley, 1989) provides a natural framework for investigating the degree to which performance assessment results are reliable or generalizable beyond the specific graders and tasks confronted by an examinee. In fact, studies of performance assessments in other contexts such as the military (e.g., Shavelson, Mayberry, Linn, & Webb, 1990) or medical licensure testing (e.g., Swanson, Norcini, & Grosso, 1987) suggests that there is likely substantial variability due to task sampling. Similarly, generalizability studies of direct writing assessments that manipulate tasks also indicate that the variance component for the sampling of tasks tends to be greater than that for the sampling of raters (Brelan, Camp, Jones, Morris, & Rock, 1987; Hieronymus & Hoover, 1986).

Research results which have suggested that task generalizability is suspect for many performance assessments is consistent with literature in learning and cognition that emphasizes the situation and context-specific nature of thinking (Greeno, 1989).
Nevertheless, the potential for limited generalizability across tasks is an issue that should be taken into account in the design of a performance assessment program. In practice, this is accomplished either by increasing the number of performance assessments for each student or by using a matrix-sampling design where different performance assessment tasks are administered to separate samples of students (Linn, 1991).

Another consideration for designers of performance assessments is that the content be consistent with the best current understanding of the field and at the same time reflective of what are judged to be aspects of quality that will stand the test of time. More important, the tasks selected to measure a given content domain should themselves be worthy of the time and efforts of students and raters. These considerations are especially important in view of the limited sampling that is likely to occur with performance-based measures (Linn, 1993).

As Collins, Hawkins, and Frederiksen (1990) have noted, if the use of performance assessments leads to gaps in content coverage, teachers and students are likely to underemphasize those parts of the content domain that are excluded from the assessment. They illustrate this issue by reference to Schoenfeld's (in press) description of a geometry teacher in New York who had been recognized for superior teaching on the basis of performance of his students on the Regents Geometry Exam. Unfortunately, the superior performance of his students was achieved by having them memorize the 12 proofs that might appear on the Regents Geometry Exam. The lack of adequate content coverage in this example clearly led not only to misleadingly high scores, but more important, to a distortion of the instruction provided.
Messick's (1989) discussion of the consequential basis of validity provides a convincing case that consequences should be a major focus of the validation of the uses and interpretations of any measure. The need to obtain evidence about consequences is especially compelling for performance-based assessments such as those envisioned by NCEST (the National Council on Education Standards and Testing), SCANS (the Secretary of Labor's Commission on Achieving Necessary Skills), and the National Education Goals Panel, because particular intended consequences are an explicit part of the assessment system's rationales (Linn, Baker, & Dunbar, 1991; Messick, 1992).

Writing Assessment

One common characteristic of many performance assessments is that examinees must produce a written document. Examples include essays, reports, summaries, brief explanations, etc. The need for writing skills in this context as well as a general perception of a decline in writing skills among US students has lead to a renewed interest in measurement of writing ability (Breland & Gaynor, 1979). Since the College Board offered its first writing examinations in 1901, two distinctly different approaches to writing skill assessment have evolved. Direct assessment requires that actual essays be written and usually such essays are read and scored independently by two or more readers. Indirect assessment, sometimes called objective assessment, requires that the examinee respond to stimuli in a multiple-choice format.

Past research on direct and indirect assessment of writing skill contain arguments supporting each. Diederich (1946) noted that unless examinees are asked to write papers that are extremely different from the writing they ordinarily have to do (or unless their papers are marked for content) "the essay is unquestionably a valid test of ability to write,
for it is an instance, a sample, of the very ability that one is attempting to measure. There is no more direct evidence of ability to write" (p. 584). Eley (1955) stressed that "an adequate essay test of writing is valid by definition; that is to say, it has face validity since it requires the candidate to perform the actual behavior which is being measured" (p. 11). Braddock, Lloyd-Jones, and Shoer (1963), objecting to multiple-choice assessment of writing skill, wrote that "not only do they [multiple-choice tests] not require the examinee to perform the actual behavior being measured--he does no actual writing; but these tests also make little or no attempt to measure the 'larger elements' of composition, even indirectly" (p. 42). Such larger elements include unity, organization, and content.

Arguing for indirect assessment, Noyes, Sale, and Stalnaker (1945) suggests that a student assessed by means of an essay is somewhat in the position of a gambler who risks all on a single throw of the dice, while a multiple-choice test allows many throws: "The good candidate who errs on a few of these [multiple-choice items] has plenty of opportunity to redeem himself; a mistake on one item does not affect any other item. In writing a theme, however, the candidate who makes a false start almost inevitably involves his whole theme in difficulties even though he may be, generally speaking, a good writer" (p. 9).

Traxler and Anderson (1935), in a study of two highly structured two-hour essays, obtained a test-retest alternate forms reliability of only .60 which tends to support the Noyes et al. position. Longer, highly structured direct assessments usually result in the highest reliabilities. Diederich (1946) states that "...essays on quite different topics, in different forms of writing, and without such rigid controls over the materials to be
included, never attain a correlation higher than .55" (p. 587). Numerous studies have reported similarly low reliabilities for direct assessment. In one such effort, Akeju (1972) concluded that the reliability of the West African General Certificate Education examinations in English composition was inadequate and suggested the use of multiple-choice tests.

Studies of the validity of direct assessments of writing skill have often revealed adequate validities despite the reliability problem. Huddleston (1954) used 20-minute essays, an objective test of composition, and a short verbal test to predict instructors’ ratings of writing skill after three months of freshman English and end-of-semester freshman English grades. Huddleston obtained validities of .43 for the objective test and the verbal test, and .34 for the essay test in predicting instructors’ ratings. When English grade was the criterion, correlations were .46 for the essay, .38 for the verbal test, and .34 for the objective test. In a second study, the essay correlated .40 with instructors’ ratings and .41 with English grades, while the corresponding correlations for the objective test were .58 and .60, and for the verbal test, .76 and .77.

Pearson (1955) studied the 1954 General Composition Test (GCT), and found that teacher ratings of student writing skills correlated .51 with GCT scores and .65 with SAT-Verbal scores. The best weighted combination of GCT and SAT-Verbal scores produced a multiple R with teacher ratings of .68, which was .03 higher than for the SAT-V alone. Checketts and Christenhsen (1974) examined the CLEP objective and essay sections and found that the essay added .06 to a correlation of .47 obtained using the objective section alone.
A general feature of most of these studies was the comparison of essay scores, based on norm-referenced holistic ratings, to total multiple-choice test scores, apparently assuming that both sets of measures tapped either a unidimensional writing ability or, at least, the same set of writing skills. However, content analysis of essay rating criteria reveal that they are often vaguely worded, but that they do reference whole-text features such as thesis, coherence, support and style, in addition to sentence-level mechanical conventions. Items on multiple-choice tests, on the other hand, often emphasize sentence-level mechanics and require few if any text-level discriminations and, obviously, no production responses (Quellmaiz, Capell, & Chou, 1982).

An example of a study which attempted to address this limitation of previous efforts was Moss, Cole, and Khampalikit's (1982) attempt to compare holistic and analytic essay ratings with a multiple-choice test’s language scores. They designed an atomistic/analytic rating scale to judge essays on the same four subscales reported for the multiple-choice measure. These included three purely sentence-level subscales and a fourth subscale "expression" referencing two sentence-level mechanical writing skills—grammar and word choice—and two vaguely defined text-level skills—organization and clarity. They found very low correlations between both the atomistic essay ratings directly matched to the purely multiple-choice mechanics subscales as well as the hybrid expression subscale. These results suggest that the essay ratings and multiple-choice item subscales were not measuring the same writing skills, that is, they were not reflecting the same underlying skill constructs.

Several other studies have attempted to address this construct validity problem by comparing direct and indirect measures of parallel text features. Smith (1978) used
domain-referenced skill specifications to design multiple-choice items analogous to essay rating criteria. She found correlations of the multiple-choice test total score with a General Impression score of .65 and with the total of analytic ratings of .61. Relationships between analogous multiple-choice and essay subscores such as organization and support, however, were much lower, ranging from .23 to .55. Taking Coffman's (1966) results on essay score reliability as a guide for assessing the impact of Smith's rating conditions on her observed correlations, the study suggests little or no improvement in direct-indirect measure agreement even when analogous measures specifications are employed (Quellmalz, Capell, & Chou, 1982).

**Writing Assessment and English as Second Language (ESL) Students**

One of the more significant recent trends in US education is the large influx of students that speak English as a Second Language (ESL) or are classified as limited English Proficient (LEP). In 1988, the U.S. Department of Commerce reported 4,710,655 immigrant arrivals to the U.S., the highest number since 1924. In California, the state with the highest growth in LEP (limited English proficient) population, 5,478,712 people reported speaking Spanish at home. Of these, 32.5% also reported speaking English "not well" or "not at all." This figure represents an 89.8% increase from 1980 to 1990. Among Asian, Pacific Island, and other language speakers in California, the percentage speaking English "not well" or "not at all' increased an astounding 100.8% between 1980 and 1990 (U.S. Department of Commerce, 1990). Added to this, there are a record number of people residing in the U.S. with F-1 (student) visas. In 1991-92, 419,585 F-1 visa students were attending U.S. colleges and universities, the highest number ever admitted (Watkins, 1992).
With the increasing diversity of the US student population, the serious writing-related problems of ESL students, and the continuing focus on assessment, researchers and educators are concerned with the fairest and most effective instructional means for ensuring a desired standard of writing for both regular students and ESL students (Benson, Deming, & Valeri-Gold, 1992). Although regular and ESL students share many of the same writing problems (e.g., sentence structure, punctuation, control over grammatical structures, rhetorical coherence, etc.) there is evidence of distinct differences between the two populations (Shaughnessy 1977; Santos 1988; Vann et al. 1988; Connors & Lunsford 1989; Kroll, 1990). For example, Benson, Deming, Denzer, and Valeri-Gold (1992) compared the written products of both regular and ESL students at a suburban two-year college. The results of the study indicated that both groups used topic sentences and preferred exposition. However, the regular students wrote longer compositions and averaged fewer errors in the construction of verb tenses, the use of prepositions, articles, and diction than did the ESL students.

The use of direct assessments of writing with ESL students raises many of the issues concerning fairness discussed by Linn(1993) relative to performance assessments in general. Are, for example, observed differences in direct writing assessments between regular and ESL students due to differences in achievement or simply differences in familiarity with the content or format of the assessment? Are direct assessments of writing as efficient at measuring mechanical writing skills as the indirect multiple-choice approach for these two distinct populations? These issues suggest the possibility of an interaction between the format of a writing assessment and the English skill level of the examinee. This study is an exploration of this issue.
Statement of Problem and Research Questions

The recent popularity of performance assessments has been accompanied by considerable research on validity and reliability criteria that these measures should satisfy. This literature has focused on the relationship of performance assessments with other (external) measures of traits of interest (e.g., supervisor or teaching ratings), as well as comparisons of performance-based and objective measures of a given trait. This latter line of study is of interest because if performance-based and objective tests yield similar information about examinees, then in many contexts the more efficient objective measures are to be preferred. Further, if these measures yield similar scores or rankings of examinees, the superior reliability of objective tests would support their use over performance-based measures.

In the present study, concern is with the similarity of results from objective (indirect) and performance-based (direct) assessments of writing in a high-stakes testing program. Do these measures yield similar results overall and/or for specific skill areas (e.g., punctuation, sentence structure, etc.) in an environment in which they carry significant consequences for examinees? Are similar constructs measured? Finally, are these results impacted by the experience of examinees with the English language?

Significance

Performance assessments hold promise as measures of traits that are of intrinsic value, and yet they pose significant challenges relative to traditional multiple-choice measures: they are less efficient than objective measures; they provide examinees with fewer tasks and increase the possibility of a ‘task’ selection bias; and they are more difficult to evaluate objectively and fairly. Despite these issues, the popularity of
performance-based measures of student achievement continues to increase and is common in high-stakes assessment programs. In the context of writing, this constitutes a return to practices of the last century. Now, as then, researchers continue to debate the similarity and dissimilarity of direct (performance based) and indirect (objective) measures of writing ability. This literature remains unclear, for example, if a student's skill in identifying punctuation problems in a written document is the same as those skills, which permit creation of a document free of punctuation errors. Further, the literature is silent on the extent to which experience with the English language impacts the differences and similarities of direct and indirect assessments of writing. This latter point is a significant issue in the US because of the dramatic increase in the number of students in US schools that speak English as a second language.

This study was designed to shed light on the similarity of direct and indirect assessments of writing skills for regular English speaking students and students that speak English as a second language. The focus of the study is a high-stakes examination, which will be used, in conjunction with other information, to determine if a student receives a high school diploma. The results of this study will have bearing on the use of these measures for regular and ESL students.
CHAPTER II

REVIEW OF RELATED LITERATURE

This Chapter reviews literature on changing educational assessments and comparing multiple-choice and performance-based assessment formats. A general definition of assessment approaches is offered. The theoretical groundwork for new assessment such as performance-based assessment is offered. The limitations of multiple-choice tests are criticized and new or alternative approaches to testing are elaborated. Finally, an integrative and critical review of criteria of comparisons between multiple-choice and performance-based assessment formats is provided.

Definitions of Assessment Approaches

Performance assessment, also known as alternative or authentic assessment is a form of testing that requires students to perform a task rather than select an answer from a ready-made list. For example, a student may be asked to explain historical events, generate scientific hypotheses, solve math problems, converse in a foreign language, or conduct research on an assigned topic. Experienced raters either teachers or other trained staff then judges the quality of the student's work based on an agreed-upon set of criteria. This new form of assessment is most widely used to directly assess writing ability based on text produced by students under test instructions (Sweet, 1992). "Performance assessment is a systematic attempt to observe and rate a person's ability to use skills and knowledge to achieve specified goals described in terms of four major attributes; assessment context, stimulus conditions or the test exercise, response mode, and scoring procedures" (Stiggins & Anderson, 1981, p.1).
Performance assessment is not new. If we go back to the time of the Greeks, we read about athletic contests such as the marathon and other endurance sports. If we look in the Bible, we see evidence of different types of performance and performance assessment. Musical, artistic, and literary accomplishments have been measured and rated for many years. Although the procedures used might have lacked refinement and psychometric elegance, they nevertheless were examples of performance assessment (Stiggins, 1981).

However, the multiple-choice item can be characterized as a stem that describes a problem, and a series of options or alternatives, each representing possible answer to the stem. Normally, one option is correct, with the remaining alternatives referred to as distractors or foils (Oosterhof, 1994). Linn (1995) defines a multiple-choice item as a problem and a list of suggested solutions. Linn (1995) further explains that the problem may be stated as a direct question or an incomplete statement and is called the stem of the item. The list of suggested solutions may include words, numbers, symbols, or phrases and are called alternatives (also called choices or options). The student is typically requested to read the stem and the list of alternatives and to select the one correct or best alternative. The correct alternative in each item is called the answer, and the remaining alternatives are called distractors (also called decoys or foils). These incorrect alternatives receive their name from their intended function to distract those students who are in doubt about the correct answer (Linn, 1995).
Sources for Performance Assessment

The Necessity for a Developmental Perspective

Over the past few decades, the various assumptions on which testing edifice was based have been gradually undermined by work in developmental, cognitive, and educational studies, and a quite different view has emerged (Gardner, 1992).

Owing to the pioneering work of Jean Piaget (1983), it is widely recognized that children are not simply miniature versions of adults. The infant or the toddler conceives of the world in a way which is internally consistent but which deviates in important particulars from a more mature conception. Here are some of the most familiar instances from the Piagetian canon: the infant does not appreciate that an object continues to exist when it has been removed from view; the toddler does not understand that material remains constant in quantity, even when its physical configuration has been altered (for example, squashing a ball of clay); the young school child is unable to reason solely from the implications of one proposition to another but instead proceeds on the basis of knowledge of concrete instances and perceived empirical regularities (Piaget, 1983).

According to Piaget's view, children pass through a number of qualitatively different stages called sensor-motor, pre-operational, concrete operations, and formal operational. A child at one stage in one area of knowledge will necessarily be at the same stage in other domains of experience (Piaget, 1983). Few investigators hold any longer to a literal version of this "structured-stage" perspective; there have been too many findings, which do not support it (Brainerd 1988; Gelman 1988). But most developmental psychologists continue to subscribe to the point of view that the world of the infant or toddler has its own peculiar structures; many developmentalists believe that there are
stage sequences within particular domains of experience (for example, language, moral judgment, understanding of physical causality); and nearly all emphasize the need to take into account the child's perspective and level of understanding (Case 1985; Feldman 1987; Fischer 1990).

Another feature of this approach is its assumption that development is neither smooth, nor unilinear, nor free of perturbations (Gardner, 1992). While details differ among theorists, most researchers believe that there may be critical or sensitive periods during which it is especially easy or especially difficult to master certain kinds of materials. Similarly, while youngsters tend to improve in most areas with age, there will be periods of more rapid growth and periods of stasis. And a minority of researchers believes that in some domains there may actually be regressions or "U-shapes" with younger children performing in a more sophisticated or integrated fashion than students in middle childhood (Strauss, 1982).

It is possible to construct measurement instruments, which reflect the developmental knowledge recently accrued. In fact, some batteries have been devised which build specifically on Piagetian or allied notions (Uzgiris and Hunt, 1966). For the most part, however, American tests have been insensitive to developmental considerations.

The Emergence of a Symbol-System Perspective

Over the past few decades, however, there has been increasing recognition of the importance in human cognition of the capacity to use various kinds of symbols and symbol systems (Gardner, Howard, and Perkins 1974; Goodman 1976; Langer 1942). Humans are deemed the creatures par excellence of communication, who garner
meanings through words, pictures, gestures, numbers, musical patterns, and a whole host of other symbolic forms. The manifestations of these symbols are public; all can observe written language, number systems, drawings, charts, gestural languages, and the like. However, the mental processes needed to manipulate such symbols must be inferred from the performances of individuals on various kinds of tasks. Unexpectedly potent support for the belief in internal symbol-manipulation has come from the invention and widespread use of computers; if these human-made machines engage in operations of symbol use and transformation, it seems ludicrous to withhold the same kinds of capacities from the humans who invented them (Newell and Simon, 1972).

Gardner (1992) points out that considerable effort has been expended in the relevant science to investigate the development of the human capacity for symbol use. It is widely (though not universally) agreed that infants do not use symbols or exhibit internal symbolic manipulation and that the emergence of symbol use during the second year of life is a major hallmark of human cognition. Thereafter, human beings rapidly acquire skill in the use of those symbols and symbol systems, which are featured in their culture. By the age of five or six most children have acquired a "first draft" knowledge of how to create and understand stories, works of music, drawings, and simple scientific explanations (Gardner, 1982).

In literate cultures, however, there is a second level of symbol use. Children must learn to utilize the invented symbol or (notational) systems of their culture, such as writing and numbers. With few exceptions, this assignment is restricted to school settings, which are relatively decontextualized. Mastering notational systems can be difficult for many students in our society, including students whose mastery of "practical
knowledge" and "first-order symbol systems" has been unproblematic. Even those students who prove facile at acquiring notational systems face a non-trivial challenge: they must mesh their newly acquired "second-order" symbolic knowledge with the earlier forms of "practical" and "first-order" symbolic knowledge they brought with them to school (Bamberger 1982; Gardner 1986, Resnick 1987).

Gardner (1992) believes that nearly all formal tests (for example, multiple-choice tests) presuppose that their users will be literate in the second-level symbol systems of the culture. These tests thus pose special difficulties for individuals whom, for whatever reason, have had difficulty in attaining second-level symbol knowledge or cannot map that knowledge onto earlier forms of mental representation. Moreover, individuals with well-developed second-level symbolic skills can often "psyche out" such tests, scoring well even when their knowledge of the subject matter which is ostensibly being assessed is modest (Gardner, 1983).

The Decomposability and the Decontextualization of Assumptions

Resnick and Resnick (1992) note that two key assumptions of standardized testing technology, which are termed as the decomposability and the decontextualization assumptions, were compatible with the routinized skill goals of the mass educational system and with the psychological theories of the first part of this century. They are, however, incompatible with thinking goals for education and with what we know today about the nature of human cognition and learning (Resnick & Resnick, 1992).

Thorndike (1922) assumed that thought could best be described as a collection of independent pieces of knowledge. Thorndike (1922) showed how the content of the elementary school arithmetic curriculum could be analyzed as a collection of "bonds"
between stimuli and responses. Thorndike proposed that the task of arithmetic instruction was to exercise all of the bonds that constitute arithmetic, rewarding correct responses and "sampling out" incorrect ones. Students who acquired all of the bonds could be said to know arithmetic completely. Students who acquired fewer bonds, or who learned them to a less reliable criterion of performance, could be said to have measurably less arithmetic knowledge.

The decomposability assumption suggests that in teaching mental competencies it will suffice to teach each of the components. It supports a notion of teaching and testing separate skills, reserving their composition into a complex performance for some indeterminate later time. For a domain such as arithmetic computation, this analysis into components, practice of the components, and testing by sampling components seems reasonable. However, for thinking skills dependent on complex knowledge (for example, understanding a written passage, writing a composition, solving a mathematics problem, or interpreting the results of a science experiment), analysis into components fails to capture the organic whole that we recognize as true competence (Resnick & Resnick, 1992).

The decomposability assumption has been seriously challenged by recent cognitive research, which recognizes that complicated skills and competencies owe their complexity not just to the number of components they engage but also to interactions among the components and heuristic for calling upon them. Complex competencies, therefore, cannot be defined just by listing all of their components (Resnick & Resnick, 1992). Information-processing theories of cognition (Anderson 1983; Newell & Simon 1972), for example, analyze cognitive performance into complexes of rules, but
performances critically depend on interactions among those rules. Each rule can be thought of as a component of the total skill, but the rules are not defined independently of one another. The "competence" of a problem solving system thus depends on how the complex of rules acts together.

Decontextualization is the second major assumption built into standardized tests. Closely linked to the decomposability assumption, decontextualization asserts that each component of a complex skill is fixed, and that it will take the same form no matter where it is used. According to this assumption, if students know how to distinguish a fact from an opinion, for example, they know how under all conditions of argument and debate, in all knowledge contexts (Wiggins, 1989). If this assumption were valid, it would be sensible to select key critical thinking skills for decontextualized practice in school. But the assumption no longer appears valid. Recent developments in the epistemology and philosophy of science (for example, Lakatos 1978; Toulmin 1972) show there is no absolute link to be drawn between fact and theory, data and interpretation. Instead, what is counted as fact depends on a complex of tools and instruments that have theories built into them and on communally accepted methods for deciding among competing assertions. Thus, not only history and literature, but also science and mathematics must be understood as interpretive domains in which knowledge and skill cannot be detached from their contexts of practice and use. Educationally this suggests that we cannot teach a skill component in one setting and expect it to be applied automatically in another. That means, in turn, that we cannot validly assess a competence in a context very different from the context in which it is practiced or used (Gardner, 1992).
An example is writing. An important part of writing a good essay is being able to edit one's own work. This fact is sometimes used to justify teaching editing and then testing it as a component of writing skill, usually on passages in which students must detect errors and choose corrections from among several given alternatives. Such exercises measure copyediting ability reasonably well, but they are poor ways to practice composition. Editing one's own work is not just a matter of detecting errors; it is also a matter of crafting phrases and sentences to convey intended meanings. Decontextualized editing exercises, whether in class or on tests, do not reveal what people do when they edit their own work. If our educational goal were to train a community of copyeditors for publishing houses, editing tests would be entirely appropriate. But if we are trying to educate people to compose a good essay or a clear memorandum, editing tests sets a false direction. Such decontextualization does violence to the kinds of abilities we seek (Resnick & Resnick, 1992).

Evidence for the Existence of Multiple Features

In recent years, there has been a resurgence of interest in the idea of a multiplicity of intelligence. Mental phenomena have been discovered that some researchers construe as evidence for mental modules--fast-operating, reflex-like, information-processing devices which seem impervious to the influence of other modules. The discovery of these modules has given rise to the view that there may be separate analytic devices involved in tasks like syntactic phrasing, tonal recognition, or facial perception (Fodor, 1983).

A second source of evidence for a multiplicity of intelligence has been the fine-grained analysis of the mental operations involved in the solution of items used in intelligence tests (Sternberg, 1977). These analyses have suggested the existence of
different components which contribute to success on any standard intellectual assessment. Individuals may differ from one another in the facility with which the different components operate, and different tasks may call upon a differential use of the various components, meta-components, and sub-components (Sternberg, 1983).

According to Gardner's six kinds of analyses (language, logical-mathematical analysis, spatial representation, musical analysis, bodily kinesthetic thinking, and intrapersonal knowledge), most formal testing (for example, multiple-choice test)—whatever the area that is allegedly being tested—engages primarily the linguistic and logical-mathematical faculties. If one has high linguistic and logical-mathematical intelligence, one is likely to do well in school and in formal testing. Poor endowment or learning in one or both of these intelligence is likely to result in poor standardized scores (Gardner, 1992).

If life consisted solely of schooling, most formal tests would serve their purpose well though last year's grades would fulfill the same predictive purposes equally well. Schooling, however, is supposed to be a preparation for life, and there is ample evidence that formal testing alone is an indifferent predictor for success once a student has left school (Jencks, 1972).

Gardner (1992) suggests assessment, which looks directly at an individual's skills in areas such as music, spatial knowledge, or interpersonal understanding, rather than looking through the "window" of linguistic and/or logical-mathematical prowess. It is the desire for modes of assessment that can detect capacities in the other intelligence, even in the face of indifferent linguistic or logical-mathematical capacities, which animates much of the applied research program (Gardner, 1992).
Recognition of Individual Differences

A consequence of the "multiple intelligence" perspective is the recognition that instead of a single dimension called intellect, on which individuals can be rank-ordered, there are vast differences among individuals in their intellectual strengths and weaknesses and also in their styles of attack in cognitive pursuits (Kagan & Kogan, 1970).

Formal tests can be an ally to the recognition of different cognitive features, but only if the tests are designed to elicit—rather than mask—these differences (Cronbach and Snow, 1977). It is particularly important that instruments used in "gatekeeping" niches (like college admissions) be designed to allow students to show their strengths and to perform optimally. Until now little effort has been made in this regard and tests are more frequently used to point up weaknesses than to designate strength (Gardner, 1992).

The Desirability of Assessing Learning in Context

When standardized tests and paradigmatic experimental designs were first introduced into non-Western cultural contexts, they led to a single result: preliterate individuals and others from non-Western societies appeared to be much less skilled and much less intelligent than Western control group. An interesting phenomenon was then discovered. Simple alterations of materials, test setting, or instructions frequently elicited dramatic improvements in performance. The "performance gap" between the subjects from another culture and the subjects from our own culture narrowed or even disappeared when familiar materials were used, when knowledgeable and linguistically fluent examiners were employed, when revised instructions were given, or when the "same" cognitive capacities were tapped in a form which made more sense within the non-Western context (Laboratory of Comparative Human Cognition, 1982).
Gardner (1992) believes assessment materials designed for one target audience cannot be transported directly to another cultural setting; there are no purely culture-fair or culture-blind materials. Every instrument reflects its origins. Formal tests that make some sense in a Western context do so because students are accustomed to learn about materials at a site removed from the habitual application of such materials; however, in unschooled or lightly schooled environments, most instruction takes place in situ, and so it only makes sense to administer assessments which are similarly in context.

Building upon this cross-cultural research, there is also an accumulation of findings about the cognitive abilities of various kinds of experts. It has been shown that experts often fail on "formal" measures of their calculating or reasoning capacities but can be shown to exhibit precisely those same skills in the course of their ordinary work—such as tailoring clothes, shopping in a supermarket, loading dairy cases onto a truck, or defending one's rights in a dispute (Lave 1980; Rogoff 1982; Scribner 1986). In such cases, it is not the person who has failed but rather the measurement instrument, which purported to document the person's level of competence.

**General Features of Performance-Based Assessment**

Newmann and Archbald (1992) propose specific criteria for performance-based assessment that help us clarify characteristics of performance-based assessment and identify key distinctions between performance-based assessment and traditional assessment (for example, multiple-choice form) in academic achievement.

Their criteria are described below:

Production of Knowledge: concerns about "persons in the diverse fields who face the primary challenge of producing, rather than reproducing, knowledge" (p. 72), about
their expressing this knowledge in the form of discourse, things, and performances; that is, through production of original conversation and writing, through repairing and building of physical objects, through artistic, musical, and athletic performance" (p. 72), about "students' setting their sights on authentic expressions of knowledge and honing their skills through guided practice in discourse, in manipulating objects, and in preparing for artistic and musical performances" (p. 72), about "the conventional curriculum which asks students identify the discourse, things, and performances that others have produced (for example, by recognizing the difference between verbs and nouns, between socialism and capitalism; by matching authors with their work; by correctly labeling rocks and body parts)" (p. 72), about "the production of knowledge that must be based upon understanding of prior knowledge, but the mere reproduction of that knowledge that does not constitute performance-based achievement" (p. 72).

Disciplined Inquiry: "a second defining feature of performance-based assessment of academic achievement is its reliance upon a particular type of cognitive work which can be summarized as disciplined inquiry. Disciplined inquiry, in turn, seems to consist of three features: use of a prior knowledge base; in-depth understanding rather than superficial awareness; and production of knowledge in integrated rather than fragmented form". (p. 73).

Prior Knowledge Base: "for new knowledge to be significant and valid, it must be based on substantive and procedural knowledge that has been accumulated through previous workers in a field who establish facts, vocabularies, concepts, theories, algorithms, and conventions for the conduct and expression of inquiry itself. The ultimate point of disciplined inquiry is to move beyond former knowledge-through criticism and..."
development of new paradigms which itself is stimulated by the foundations of prior knowledge. Most of the cognitive work of school consists in transmitting prior knowledge to students and asking them to reproduce it, rather than helping them to use it to produce knowledge" (p.73).

In-Depth Understanding: concerns about "developing in-depth understanding of a problem, rather than only passing familiarity with or exposure to pieces of knowledge, about facilitating complex understanding on relatively limited, special problems", about "such detailed understanding which is particularly necessary for production of new knowledge", and about "many of the cognitive tasks of school that ask students to show only superficial awareness of a vast number of topics" (p.73).

Integration: the extent to which one "must assemble and interpret information, formulate ideas, and make critiques which cannot be easily retrieved from the existing knowledge base--all of which require the ability to organize, synthesize, and integrate information in new ways" (p.73), the extent to which success in such tasks "is unlikely unless students learn to look for, to test, and create relationships among pieces of knowledge that otherwise appear unconnected" (p.73), the extent to which scientific theories, literary and artistic masterpieces, architectural and mechanical designs, musical compositions, or philosophical arguments "must ultimately be encountered as wholes, not as collections of knowledge fragments" (p.73), the extent to which too often tests of achievement "ask the student only to show recollection of unrelated knowledge fragments: definitions of terms; short descriptive identifications of people, things, events; or numerical solutions to problems" (p.74), the extent to which students "demonstrate proficiency by giving short responses, as in a TV quiz show, where answers bear little
relation to one another", and the extent to which success on these exercises "may contribute to production of integrated knowledge, but cannot be considered an indicator of it" (p.74).

Value Beyond Evaluation: concerns about performance-based assessment that "has aesthetic, utilitarian, or personal value apart from documenting the competence of the learner" (p.74), about achievements that "have special value which is missing in tasks contrived only for the purpose of assessing knowledge (such as spelling quizzes, laboratory exercises, or typical final exams)" (p.74).

Newmann and Archbald (1992) summarize that performance-based assessment requires students to engage in disciplined inquiry to produce knowledge that has value in their lives beyond simply proving their competence in school. Mastery of this sort is unlikely to be demonstrated in familiar testing and grading exercises. Instead, it is more often expressed in the completion of long-term projects, which result in discourse, things, and performances of interest to students, their peers and the public at large.

Newmann and Archbald (1992) have recognized that two compelling things are related to the problems of conventional testing. Newmann and Archbald stated, "First, participation in performance-based task is more likely to motivate students and to sustain the hard work that learning requires. Because performance-based work has value beyond the demonstration of competence in school and because it permits more comprehensive use of the mind, students will have a greater stake in performance-based achievement. Second, performance-based academic challenges are more likely to cultivate the kind of higher-order thinking and problem-solving capacities useful both to individuals and to the society. The mastery gained in school is likely to transfer more readily to life beyond
school, which increases the efficiency of our investment in schooling" (Newmann & Archbald, 1992, P. 245).

Criteria of Comparisons between Multiple-Choice and Performance-Based Formats

Success Rate

Research on the comparability of the construct validity of multiple-choice and performance-based formats has been somewhat sparse and varied in terms of conclusions. An early experiment was conducted in which the results of multiple-choice and open-ended techniques were compared on a test of vocabulary (Test E of the Self-judging Vocabulary Scale) (Heim & Watts, 1967). The subjects were two groups, each consisting of fifty naval ratings. Group M took the vocabulary test in its multiple-choice form immediately followed by the same test in open-ended form. For Group O, the order was reversed. Heim and Watts (1967) found significant differences between multiple-choice and open-ended scores. First, the multiple-choice technique provided a significantly easier task than did open-ended answering, when groups matched on an intelligence test took the same vocabulary test in the two forms. Second, the order in which the two forms of vocabulary test were taken was found to influence (a) the subjects' responses and (b) the association between intelligence test and vocabulary test scores. Third, the subjects' scores on both forms of the test proved to be fairly closely related to their self-estimates of word-familiarity, despite the fact that the groups tested were not very highly literate. Heim and Watts (1967) indicated that tests employing different formats cannot be expected to have the same means, standard deviations, and correlations with criterion variables. Some of these differences can be assumed to be due to changes in the scale of
measurement and amount of error variance associated with each format. Thus, the results of this early experiment do not necessarily imply lack of construct equivalence.

White and Carcelli (1982) investigated the effect on children's test scores of different item formats used in standardized mathematics achievement tests. Second grade students completed a mathematics computation test using eight different formats derived from five standardized achievement tests. Identical content, taken from most frequently occurring content in the standardized achievement tests examined, was tested with each format. Differences in test scores between types of formats were statistically significant. The results of this study indicate that children score very differently on mathematics computation items depending on the format in which the test items are presented. This means that conclusions about how well children have mastered mathematical content will depend in part upon the format of the particular standardized test which is used (White & Carcelli, 1982).

White and Carcelli (1982) indicated that the students obtained significantly lower scores on problems that were verbally administered. It may be that children received lower test scores with verbally administered problems because the test stimuli were transient and required the children to remember the type of problem and the value of the numbers being computed while at the same time determining the correct answer.

Children scored higher on items that used multiple-choice response formats than on items that used open-ended response formats. With a multiple-choice format, children needed only recognize the correct response and mark it. However, with the open-ended format, children needed to retrieve the correct response from memory before answering.
Apparently, retrieving a correct answer from memory is more difficult than simply recognizing the correct answer (White & Carcelli, 1982).

**The Traits Measured By the Test**

Traub and Fisher (1977) used verbal-comprehension and mathematical-reasoning tests in a similar study. The stems of the items for the two formats were identical. The free-response versions were given first, to eliminate learning from multiple-choice cues, and the multiple-choice version was given two weeks later. Traub and Fisher (1977) recognized these problems and employed methods that equated scale parameters and error variances on three response formats for both verbal and quantitative measures. Two of these formats were multiple-choice and constructed-response. Using confirmatory factor analysis (CFA), Traub and Fisher (1977) found little evidence of a format effect for the mathematical reasoning items, and only weak evidence that the free-response and multiple-choice items were measuring different constructs for verbal comprehension items.

Similarly, Ward, Frederiksen, and Carlson (1980), also using CFA, compared machine-scored and constructed response forms of a test of ability to formulate scientific hypotheses. It was found that a knowledge factor, based on the GRE Advanced Psychology Test and verbal and reasoning factors had quite similar patterns of correlation with various scores for both forms, especially for scores representing the quality of the hypotheses. The most striking difference had to do with the role of ideational fluency (Guilford, 1967, 1982), which we interpret as skill in broadly searching long-term memory for relevant ideas. None of the scores from the multiple-choice form had appreciable correlation with fluency measures. For the free-response form, however,
scores reflecting number of ideas, number of unusual ideas, and the number of ideas that are both unusual and of high quality, all had substantial correlations with fluency, even though they were less reliable than the comparable scores based on multiple-choice responses. Only the free-responses form appeared to require a broad search for relevant ideas and information stored in memory. Thus, format does make a difference for at least one problem-solving test.

Ward (1982) investigated three types of verbal items—antonyms, sentence completion, and analogies—each measured using several formats, including multiple-choice and free-response forms. Ward (1982) concluded that for verbal aptitude items, various item formats produce much the same information and are essentially equivalent in terms of both the technical adequacy of the resulting measures and the construct interpretations of the resulting scores.

A study by Godshalk, Swineford, and Coffman (1966), using both essay and multiple-choice assessments of writing for 11th and 12th grade students, concluded that (1) "When objective questions specifically designed to measure writing skills are evaluated against a reliable criterion of writing skills, they prove highly valid." and (2) The most efficient predictor of a reliable direct measure of writing ability is one which includes both essay questions and objective questions" (p.41).

Although research has cited the predictive or concurrent validity of indirect writing measures (Breland, 1977; Coffman, 1971), these measures are regarded by a number of researchers to be weak in content, construct, and "ecological" validity (Braddock, Lloyd-Jones, & Schoer, 1963; Cooper & Odell, 1977). In essence, the critics of indirect assessment of writing, although they do concede that these measures probably
adequately measure comprehension and editing abilities, claim that the tests make little or no attempt to measure unity, content, or organization because the examinee does no actual writing (Ackerman & Smith, 1988).

There is empirical evidence to suggest that indirect and direct measures of writing ability may, in fact, be measuring different types of abilities. Most attempts to investigate the relationship between scores resulting from varieties of direct and indirect writing assessment methodologies have been primarily correlational in nature, and have focused on rather limited definitions of the two assessment strategies. Several studies have produced results that show moderate correlations (ranging from .3 and to .6) between standardized tests of language skills and holistically scored essays at both the high school and college levels (Breland, Conlon, & Rogosa, 1976; Breland & Gaynor, 1979; Coffman, 1966; Hogan & Mishler, 1980; Moss, Cole, & Khampalikit, 1982). One general conclusion that can be drawn from the results of these studies is that the two assessment methods may be assessing dissimilar skills and therefore are not of equal value in evaluating skills thought to comprise writing proficiency.

In addition, Ackerman and Smith (1988) investigated the similarity of information that is provided by direct and indirect methods of writing assessment. The skills required by each of these techniques provide a framework for a cognitive model of writing skills from which these procedures can be compared. A sample of 219 10th-grade students participated in the experiment. Students were randomly selected from traditional English classes. Three instruments were used in the study to elicit varying degrees of procedural knowledge on the part of the students. The first was a multiple-choice standardized achievement test (language and writing). A second measure was an essay task. The essay
prompt asked students to "choose one or two ways in which you feel that television is of benefit to individuals, to families, or even to all of society. Tell exactly what you consider the benefits of television to be and tell how television can improve or help people". The third instrument was a free-response version of the standardized test. This instrument was identical to the standardized test except that it required the student to generate the correct answer instead of selecting the correct alternative.

Results of this study suggest, at least in the areas of writing assessment, that the construct being measured is a function of the format of the test. Scores obtained from direct and indirect methods of writing assessment provide different information. Specially, the skill of generating topic knowledge is more accurately assessed with an essay task. The CFA (Confirm Factorial Analysis) procedures suggest that the free-response format measured the ability to organize coherent paragraphs much better than did the multiple-choice format. Ackerman and Smith (1988) note that essay scores are almost totally dominated by these tasks. That is, variance in the essay score structure is heavily dominated by higher-order generation components such as paragraph development and paragraph structure. The moderate correlations observed between direct and indirect assessment procedures can be explained by the fact that direct methods contain more types of skills than indirect.

Retention

In a study by Duchastel (1981), the relative effects on retention of different tests administered immediately after the learning of a brief text were compared. Learning was followed by either a short-answer test, a multiple-choice test, or no test at all on the passage. Retention two weeks later was superior for the group that had received the short-
answer test when compared to each of the other two groups. The interpretation advanced for this finding was that the immediate short-answer test had enhanced consolidations by initially requiring a greater degree of mental review than the multiple-choice test. Questions on a short-answer test cue the student as to what to search for but require him or her to actively retrieve that information from memory. Multiple-choice questions, on the other hand, reduce the degree of memory search required by providing the student with the correct answer as one of the alternate choices offered in the item (Duchastel, 1981).

The psychological processes underlying the type of response involved in short-answer and multiple-choice questions are different although not yet satisfactorily understood (Brown, 1976). In one case, memory search is at the core of the process; in the other case, discrimination plays the important role, with memory search thought to be secondary. The contention advanced here is that it is memory search itself which consolidates learning and the consequence that a short-answer test given immediately after learning will result in a greater testing effect (a greater retention of the material later on) than an initial multiple-choice test (Duchastel, 1981).

Kumar, Rabinsky, and Pandey (1979) conducted a study in test mode, test instruction, and retention. Sixty ninth-grade students were asked to read a passage for a test to be taken after 24 hours. Equal numbers of subjects were randomly instructed to anticipate a recall test, a multiple-choice test, and a retention test. Half the subjects in each condition received randomly either a recall or a recognition test (multiple-choice test) regardless of what they were led to anticipate prior to reading the passage. Subjects did better on the recognition tests (multiple-choice tests) as compared to the recall tests.
The effect due to test instruction given to anticipate different types of test was not significant. The interaction between actual test mode and test instruction given was not significant, indicating that subjects tested in the anticipated mode performed as well as those tested in unanticipated mode.

Kumar, Rabinsky, and Pandey (1979) note that consideration of study times favors the interpretation that subject processing strategies at the storage time were not influenced by the test mode instructions. It was found that item difficulty correlations between conditions involving the same test mode were high and positive (regardless of instruction), suggesting encoding and retrieval similarity between those conditions. The fact that intercorrelations between recall and recognition conditions were quite low (regardless of instructions) suggests that differences in the performance of the recall and recognition tests are more a function of the nature of the test mode rather than of differential encoding of information during the learning stage (Kumar, Rabinsky, & Pandey, 1979).

**Strategy Choice**

Freund, Brelsford, and Atkinson (1969) employed an anticipatory paired associate learning task using three-letter nonsense syllables (consonant-vowel-consonant pronounceable trigames of approximately 20 to 50 % association values) and found that the differences between recall and recognition scores were not related to subject’s knowledge, at time of study, of the mode of the test to be employed. Rather, they attributed the differences between recall and recognition to differences in retrieval processes. In another study, Tversky (1973), using pictures of common objects and words as stimuli, found that subjects' performance was higher on both the recall and recognition
tests when they were tested by the anticipated method than when they were tested by the unanticipated method. In other words, although subjects appeared to process differently in anticipation of a specific test mode, there did not seem to be any advantage in terms of performance by preparing for a recall test over a recognition test.

Loftus (1971) compared recognition and recall in a continuous memory task. Differences between recall and recognition performance may be due at least in part to differences in the way information is stored. Loftus (1971) explored this possibility in a paired-associate learning task by varying student's knowledge at the time of study of how he would be tested on a particular stimulus-response pair. It was found that when student knows how he/she would be tested, his/her performance was better on recall but worse on recognition than when he/she did not know how he/she would be tested. These results were interpreted as support for the assertion that differences in storage processes partially account for recall-recognition performance differences. A model which postulates a distinction between short-term and long-term memory provided an excellent fit to the data and suggested possible storage strategies for recall and recognition (Loftus, 1971).

Siegler (1988) examined individual differences in children's strategy choices. In particular, the research focused on consistencies in 6-year-olds' strategy choices on three tasks: addition, substraction, and word identification (reading). Differences were present along 2 dimensions: knowledge of problems and stringency of thresholds for stating retrieved answers. Siegler (1988) indicated that children could be classified into 3 groups: good students, not-so-good students, and perfectionists. Perfectionists were children who had good knowledge of problems and set very high thresholds for stating retrieved answers, good students also had good knowledge of problems but set lower thresholds,
and not-so-good students had less good knowledge of problems and set low thresholds. It was found that the groups differed in standardized achievement test performance 4 months after the experiment in ways consistent with the experimental analysis. The pattern of individual differences was similar in 2 experiments with different samples of children, problems and different methods for assessing strategy use. The results indicated how detailed cognitive models (performance-based assessment) can contribute to understanding of individual differences (Siegler, 1988).

**Diagnostic Information**

Most of these studies used tests of reading comprehension or vocabulary, domains in which the cognitive mechanisms underlying recall versus recognition are applicable. One question that arises, however, is whether the same distinction holds for multiple-choice and open-ended items in procedural tasks such as arithmetic operations (Birenbaum & Tatsuoka, 1987).


Birenbaum and Tatsuoka (1983) believe that the modified rule of operation (or algorithms) sometimes happen to be incorrect as a result of various misconceptions or "bugs" developed by the students during the learning process. However, some of those incorrect rules can occasionally yield correct answers. Brienbaum and Tatsuoka (1983) administered a free response test in signed numbers to students and scored them twice; once according to the conventional method that credits all right answers, while the second
scoring system was based on the underlying algorithm, thus giving no credit to right answers presumably derived by an incorrect algorithm. Both scoring systems were then compared with respect to their factorial structure and reliability.

The results of this study (Brienbaum & Tatsuoka, 1983) indicate that in achievement data of the problem solving type, where a specific subject matter area is being tested, the factorial structure of the data is highly affected by the existence of different algorithms underlying the student response of the data. The fact that students can sometimes get right answers by following a wrong rule is reflected in the psychometric properties of the test. It indicates that when the conventional scoring system is used, it results in negative correlations among some items, and increased dimensionality. When a scoring system that takes into consideration the process rather than relying solely on the outcome of the cognitive process is used, significant gains in the psychometric properties of the test can be obtained. Moreover, the variation in the amount of modifications made in the different tasks is not highly related to the amount of changes that occur in the task-total correlations as a result of the modifications. One cannot substitute the search for algorithms with the conventional discrimination index of the classical test theory (Birenbaum & Tatsuoka, 1983).

Birenbaum and Shaw (1985) studied task specification and its relationship to test results. A task specification chart which integrates the content facets and procedural steps of a set of tasks was constructed, and its efficacy in accounting for variance in item difficulties and in diagnosing students' errors was examined. Two paper-and-pencil tests, one in addition and the other in subtraction of fractions, were designed to include all elements presented in the task specification chart (Klein et al., 1981; Tatsuoka, 1984).
Both tests were of the constructed response type. The tests were administered to two different samples. Sample I consisted of 595 students in the 7th, 8th, and 9th grades in two junior high schools in two Midwestern towns in the United States. Sample II consisted of 241 7th graders from a junior high school in an average size town in Israel.

Birenbaum and Shaw (1985) focused on the effect of a task specification chart and its implications for testing and instruction. The results indicate that item characteristics representing different task components can account for most of the variance in item difficulties on tests. The relationships among the components specified in the chart have important implications for testing as well as for instruction. Using a task specification chart, the teacher can design instruction to match the difficulty of the task components. The chart also provides a basis for describing students' levels of achievement in terms of components not mastered (Birenbaum & Shaw, 1985).

The typology of errors identified in this study proves to be a useful tool for diagnosis. The classification of errors according to the task components enables the teacher to evaluate instruction and redesign it accordingly. The analysis of a student's response patterns in terms of the underlying rules of operation enables the teacher to prescribe individualized remediation to address the misconceptions reflected in that student's responses. This kind of information is above and beyond the kind provided by the traditional method of scoring a test, that is, for the total number of correct responses (Birenbaum & Tatsuoka, 1982; Tatsuoka & Tatsuoka, 1983; Birenbaum & Shaw, 1985).

Moreover, Birenbaum & Tatsuoka (1987) examined the effect of response format: open-ended versus multiple-choice on the diagnosis of examinee misconceptions in a procedural task. A test in fraction addition arithmetic was administered to 285 eighth-
grade students, 148 of whom responded to the open-ended version of the test and 137 to
the multiple-choice version. The two datasets were compared with respect to the
underlying structure of the test, the number of different error types, and the diagnosed
sources of misconception reflected in the response patterns.

Birenbaum & Tatsuoka (1987) indicated considerable differences between the
two formats, with more favorable results for the open-ended format. The underlying
favorable structure, as examined by smallest space analysis, seemed clearer in the open-ended dataset, where the configuration of the items in the two-dimensional space clearly indicated two clusters: one of items with like denominators and the other of items with unlike denominators. The item configuration for the multiple-choice dataset, on the other hand, seemed quite diffuse, with no distinct separation between the different item types.

The results of the error analysis provided an even clearer distinction between the
two response formats. Although the two groups did not differ in the ability to solve
fraction addition problems, the multiple-choice dataset included a significantly larger
number of different error types than the open-ended dataset. These results indicate that
the open-ended group can be better diagnosed with respect to bugs or sources of
misconception underlying the response patterns, and that students who have not mastered
the task tend to be less consistent in applying their rules of operation for solving
procedural tasks when faced with a multiple-choice than with an open-ended one
(Birenbaum & Tatsuoka, 1987).

Birenbaum and Tatsuoka (1987) imply that multiple-choice tests, though
considerably easier to score, may not provide the appropriate information for identifying
students' misconceptions with respect to the given subject matter. The open-ended format seems more appropriate for this purpose.

Nevertheless, Written examinations that include multiple-choice, true-false, matching, word-completion and short-answer test items are heavily used for student assessment in the clinical years of medical school and for physician assessment in licensure or specialty board examinations. Such examinations are used despite a general recognition among educators that these test formats assess only the ability to recognize or recall facts and concepts and that performance on these tests does not correlate with clinical performance (Wingard & Williamson, 1973). The persistent, universal overuse of these limited, unidimensional test items for the assessment of clinical capability may be due in part to the ease with which they can be employed and scored, often by people other than the busy clinical teacher (Barrows, Williams, & Moy, 1987).

Barrows, Williams, and Moy (1987) have selected a clinical performance test that is similar to the written test item as a particular assessment tool which does not vary between students, and is convenient in use. The simulated patient provides such a format. Using the simulated patient in a standardized task and collecting data on an examinee's performance through the use of conventional procedures (such as checklists, rating scales, and short-answer essay questions) permit the design and compilation of a comprehensive, performance-based assessment of clinical competency.

A performance-based assessment of the clinical competence of a senior class of 72 students using simulated patients was designed and first administered in January 1986 at Southern Illinois University School of Medicine. Barrows, Williams, Moy (1987) note that the performance-based assessment of clinical competency offered the same
advantages of written examinations but goes far beyond knowledge recall and recognition and minimizes the disadvantages of faculty ratings as a measure of performance. The examination also allowed a direct comparison of clinical performance among students in the same class and between classes. The results of the first administration showed that such a comprehensive assessment is feasible and not expensive, especially in the light of the knowledge gained about students' overall capacity (Barrows, Williams, & Moy, 1987).

In a 1984 conference of medical school deans, senior faculty members, and educators, participants felt that such tests of student performance on a national basis could provide a much more valid assessment of clinical competence and curriculum effectiveness than was available through written tests (Barrows & Peters, 1984).

Writing Assessment

Language Arts is a major area of instruction in the nation's schools. This area focuses on the structure and use of the English language, and serves as a foundation for writing skills. Whether or not schools are succeeding in training students to write adequately is currently a matter of great controversy and concern. The assessment of written language skills at all levels of educational achievement, which is related to this concern, has received much attention.

Moss, Cole, and Khampalikit (1982) explained procedures to assess written language skills that are typically divided into two dichotomous categories; direct procedures and indirect procedures. Direct assessments require a sample of writing from an examinee and a response from a human evaluator; they are generally thought of as "essay tests." Indirect assessments typically involve a series of multiple-choice, machine-
scorable items, and are generally called objective tests. The standard line of inquiry into writing assessment procedures compares the performance of student writers on essay tests and objective tests.

Moderate correlations between these two assessment procedures have been repeatedly found with college-level or near-college level students. For example, Breland, Colon, and Rogosa (1976) found a correlation of .42 for a group of 96 college freshmen between scores on the College Board Test of Standard Written English, a 50-item multiple-choice test, and a 20-minute essay scored on the four-point scale with two independent ratings. A second study by Breland and Gaynor (1979) reported correlations for college freshmen of .63 (n= 819), .63 (n= 895), and .58 (n = 517) between the College Board Test of Standard Written English scores and essays scored on a six-point scale by two raters. The correlation involving the sum of the three essay scores with the sum of the three scores of the College Board Test of Standard Written English was .76.

A similar study conducted by Huntley, Schmeiser, and Stiggins (1979) compared scores obtained from 3 samples of approximately 50 college students on the American College Testing Program's English Usage Test (a 75-item, multiple-choice test), and the sum of three essays, each of which was scored by two raters. The correlations obtained were .43, .50, and .67. In a now classic study of the performance of 646 junior and senior high school students, Godshalk, Swineford, and Coffman (1966) reported correlations of from .46 to .71 between the sum of 5 essay scores, each score produced by 5 independent ratings and various subtests of the College Board English Composition Test. They also found correlations ranging from .72 to .75 between the composite essay score and a
score produced by various groupings of three subtests of the College Board English Composition Test.

A study by Hogan and Mishler (1980) extended this inquiry to the elementary school level. Hogan and Mishler gave the Metropolitan Achievement Test-Language Instructional Tests (MAT-LIT) and 20-minute essay tests to approximately 140 children in Grade 3 and 160 children in Grade 8. Approximately 60 students at each grade level wrote a second essay. For the total group, the correlation between the MAT-LIT and the essay, which was scored independently by two raters with a third rater to resolve discrepancies, were .68 and .65 for Grades 3 and 8 respectively. For the subgroups completing a second essay, the correlations were .75 and .81. These authors also compared scores from another sample of children of the same age on the MAT-Language Survey Battery and a free writing sample; they found similar levels of correlation. Hogan and Mishler (1980) concluded:

Correlations between objective tests of language skills and performance in a free-writing situation at the elementary school level are of the same general magnitude as at the college entrance level.... In addition, we noted no systematic difference between grades 3 and 8 in the degree of relationship. The similarity of the relationships from grade 3 to grade 8, to the college entrance level is rather remarkable in the light of the great differences in developmental levels covered in this range. (pp. 225-226).

Moss, Cole, and Khampalikit (1982) expand this line of research in two major ways. First, they believe that the generally used dichotomous distinction between direct and indirect language assessments is oversimplified. Moss, Cole, and Khampalikit (1982)
argue that written language assessments can be seen to vary along a number of relevant dimensions that should be examined in more detail if the relationship between so-called direct and indirect assessments at any age level is to be more fully understood. They used the three dimensional framework for classifying written language assessment procedures (task structure, scoring method, and level of evaluation). Second, they examine the pattern of the relationships among these three assessment procedures at various developmental levels in elementary and secondary schools. Moss, Cole, Khampalikit (1982) found that very low correlations between both the atomistic essay ratings directly matched to the purely multiple-choice mechanics subscales as well as the hybrid expression subscale. These results suggest that the essay ratings and multiple-choice item subscales were not measuring the same writing skills.

The results of this study raise the possibility that the correlations may be lower at lower grade levels although moderate correlations between objective and essay scores have been demonstrated at the college-level or near-college-level (Moss, Cole, & Khampalikit, 1982).

Although moderate correlations between direct and indirect methods of writing assessment are routinely reported in the literature, little objective evidence exists pertaining to why these correlations are not larger, or what unique information is provided by each of the two assessment methods. Although it has been argued that the most efficient and objective method of writing skill assessment is through the use of indirect methods (Breland, 1977), language arts experts and cognitive theorists point out that such methods fail to effectively assess many important aspects of the writing process (Braddock et al., 1963).
Ackerman and Smith (1988) investigated the similarity of information that is provided by direct and indirect methods of writing assessment. They provided information regarding the unique skills and/or abilities measured by each approach using a conceptual framework for the writing process proposed by Hays and Flower (1980).

The difference between writing an essay and answering a series of multiple-choice items can be considered in terms of the types and numbers of procedural skills that each task demands. The objective test item model requires only the editing and reading skills (i.e., primarily declarative knowledge) to select an appropriate solution. The writing task demands the procedures of setting goals, generating information, organizing this information, imposing a grammatical framework on it, and then reviewing it for possible errors in meaning or structure; thus the task requires both declarative and procedural knowledge (Hayes & Flower, 1970).

Ackerman and Smith (1988) randomly selected 219 10th-grade students from traditional English classes. Three instruments were used in the study to elicit varying degrees of procedural knowledge on the part of the students. The results of this study suggest that the construct being measured is a function of the format of the test. Scores obtained from direct and indirect methods of writing assessment provide different information. Specially, the skill of generating topic knowledge is more accurately assessed with an essay task. The confirmatory factor analysis procedures suggest that the free-response format measured the ability to organize coherent paragraphs much better than did the multiple-choice format. Variance in the essay score structure is heavily dominated by higher-order generation components such as paragraph development and paragraph structure. The moderate correlations previously observed between direct and
indirect assessment procedures can be explained by the fact that direct methods contain more types of skills than indirect. These results suggest that the overlap may not be totally eclipsed; rather, indirect methods can be characterized by declarative components of text generation (Ackerman & Smith, 1988). The findings of this study indicate that direct methods provide a better measure of procedural-type writing skills, and that indirect methods should be preferred for measuring declarative-type-writing skills.

Summary

This chapter reviews some of the evidence on which this shifting psychological conception has been based, highlights the principal features of this perspective, and indicates where it may clash with standard views of testing because alternative picture of assessment builds on the newly emerging picture of human development.

The results of these studies are quite inconclusive. Some indicate an advantage for the performance-based assessment, while others show no difference between the two formats.

The research suggests some differences in success rate can be attributed to the different abilities required to correctly answer a problem in the two formats. What a student appears to have mastered in a subject area is substantially influenced by the format of the particular test used. Some of these differences can be assumed to be due to changes in the scale of measurement and amount of error variance associated with each format. It does not imply lack of construct equivalence.

The studies of format differences indicate that if we compare existing multiple-choice tests with their free-answer counterparts, format has little influence on what construct is measured. However, when we begin with existing free-response tests
designed to measure more complex cognitive problem-solving skills, different results are found. Translations of such tests into a multiple-choice format do not measure the same construct, as do the free-response forms.

Some investigations are generally consistent in showing that testing increases retention of the material tested and that the effects are quite specific to what was tested. There is some evidence that short-answer or completion tests may be more conducive to long-term retention. The differences are not dramatic, possibly because the researchers usually tried to make the tests as much alike as possible in all respects except format, even to the point of using identical item stems. Thus they have ruled out any effects that might be attributable to the influence of format on the content of the items. Some studies indicate that recognition test scores are significantly higher than the recall test scores and are more a function of the nature of the test mode (i.e., problem at retrieval stage) rather than of differential encoding of information during the learning stage (Kumar, Rabinsky, & Pandey, 1979).

The results of studies in recognition vs. recall indicate differences between recognition and recall do not depend on whether or not the subject knew, at time of study, the mode of test be employed. These results are interpreted as support for the assertion that differences in retrieval processes are sufficient to account for differences in recognition and recall (Freund, Brelsford, & Atkinson, 1969). Other studies suggest that subjects' performance is higher on both the recall and recognition tests when they are tested by the anticipated method than when they are tested by the unanticipated method (Tversky, 1973). But Siegler (1988) indicates that multiple-choice testing generally does not provide information appropriate to assess a student's higher-order cognitive processes.
such as strategy choice. Siegler used a previously formulated cognitive model to classify three groups of individual differences in first graders' strategy choices in addition, subtraction, and reading.

Brienbaum and Tatsuoka (1987) evaluated the effect of the response format (MC vs. OE) on the rules of operation underlying examinees' response patterns in fraction-addition arithmetic items and indicate that there are considerable differences between the two formats, with more favorable results for the open-ended format in terms of the underlying structure of the test, the number of different error types, and the diagnosed sources of misconception reflected in the response patterns. Barrows, Williams, and Moy (1987) have compared performance-based assessment and multiple-choice testing in terms of predictive power and diagnostic information. The study indicates that performance-based assessment provides better indicators of student field-based performance.

Moderate correlations between standardized tests of language skills and holistically scored essays exist at elementary, high school and college levels. It indicates that direct and indirect measures of writing ability may assess dissimilar skills and therefore are not of equal value in evaluating writing skills. Ackerman and Smith (1988) used confirmatory factor analysis to examine differences in student writing abilities and found that direct method measures higher-order generation skills in paragraph development, paragraph structure, and contain more types of skills than indirect method.

The linguistic and cultural diversity among students in American schools is greater now than at any time since the early decades of this century. More than one-fifth of school-age children and youth come from language minority families in which...
languages other than English are spoken. For many of these students, English is not their first language and they enter school with limited English proficiency. During the 1980s, the number of LEP students grew two-and-a-half times faster than regular school enrollment (Lucas, 1993).

Although LEP students represent more than 100 different language backgrounds, Spanish is the native language of 65 to 70 percent of all LEP students, while 10 to 15 percent speak one of several Asian languages. Over 40 percent of students with limited English proficiency are immigrants. When they first enter American schools, LEP students vary greatly in age, mastery of English, literacy in their native language, academic preparation, and familiarity with American culture. Some immigrants students have had excellent education in their home country, while the schooling of others has been of poor quality, sporadic, or interrupted by war or other social crises (Anderson, 1994).

Nearly all LEP and other language minority students are members of ethnic and racial minority groups and most are poor. Their neighborhoods are likely to be segregated and beset with multiple problems—inadequate health, social, and cultural services; insufficient employment opportunities, crime, drugs, and gang activity. Their families are likely to suffer the stresses of poverty and to worry about their children’s safety in a dangerous environment and about their future with few positive prospects (Minicucci & Olsen, 1992).

Research studies have suggested that language minority students in general take fewer academic courses, lag significantly behind in writing, science, and mathematics, and have much higher dropouts rates than white, native-English-speaking students. At the
secondary school level, LEP students are unlikely to be given access to a full academic program taught in their native language or with special assistance (Mnicucci & Olsen, 1992).

Despite these findings, the current educational reform movement, which aims to improve the academic achievement of all students may be ignoring the needs of LEP students. There appears to be a large gap between education reform efforts for native-English-speaking students and the kinds of programs generally available to LEP students. Too often, schools undergoing restructuring fail to include LEP students in their attempts to reform the educational program (Nelson, 1996).

Thus, the challenge for schools with LEP students is to integrate the tenets of education reform with knowledge about learning in a second linguistic and cultural environment. LEP students must first of all have access to challenging curricula in language arts, mathematics, science, and other academic subjects. Simultaneously, schools must deliver a high quality academic program, which requires teachers and administrators to select appropriate instructional methods and redesign the school structure to enhance the achievement of LEP students (Anderson, 1994).
CHAPTER III
METHODOLOGY

Overview

This chapter begins with an overview of the assessment instruments and data used in this study. Following this, the sampling procedures utilized will be discussed. Finally, the analyses undertaken in this effort will be described.

This study focused on achievement levels on the subskill areas (sentence structure, usage, etc.) measured in the multiple-choice Language Arts Test and the performance-based Written Composition Test included as parts of the Graduation Exit Examination (GEE) component of the Louisiana Educational Assessment Program (LEAP). A multivariate analysis strategy was used to study group effects (regular Vs LEP students) on performance across the two assessment modalities. This study also used exploratory factor analysis and confirmatory factor analysis to test for invariance of the latent structure of the multiple-choice and Written Composition tests across the two groups.

Instruments and Data

The GEE measures curricula-based proficiencies in English Language Arts, Mathematics, Written Composition, Science, and Social Studies. The five GEE components are administered each April to all public high school students in Louisiana--to receive a high school diploma, public school students must pass all five components. Students who do not achieve the performance standards on examination components have retake opportunities in the fall and the spring of each year; seniors scheduled for
graduation have two additional retake opportunities in the winter and summer. Students who cannot pass all five components of the examination will not be awarded a high school diploma.

This study focused on the English Language Arts and Written Composition portions of the GEE. The English Language Arts test is a multiple-choice (MC) criterion-referenced measure that reflects proficiency in eight subskills: vocabulary, comprehension of details, comprehension of main idea, comprehension of critical reading, writing mechanics, word usage, sentence structure, and study skills. There are a total of 57 multiple-choice items on the test. The Written Composition Test consists of a writing prompt administered annually to Grade 10 students. The writing prompt is scored on five dimensions that comprise the general skill complex of written composition: responsiveness to assignment, support/elaboration/organization, sentence formation, usage, and mechanics. The scoring of each dimension is scaled along five levels: (0) Non-Scorable; (1) Little or no control, (2) Limited Control, (3) Moderate Control, and (4) Consistent Control. The skill areas sentence structure, word usage, and writing mechanics are common to both tests.

Due to the high stakes nature of the program, the Louisiana Department of Education has conducted extensive reliability and validity studies of both the English Language Arts and the Written Composition tests. Internal consistency, Kappa, and interrater agreement indices for these measures are all above the 0.80 range. Similarly, content validity studies have shown that these measures provide representative samples of the intended portion of the state's curriculum (Lang, 1995).
Sample

The target population for this study consisted of all grade 10 regular education and Limited English proficient (LEP) students enrolled in public schools in Louisiana during the 1995-1996 school year. The accessible population consisted of the segment of this population which participated in the GEE. This proportion is approximately 99% because all public school students in Louisiana must take and pass the GEE before they can be awarded a high school diploma.

A total of 49,373 grade 10 regular education students and 386 grade 10 limited English proficient students took the GEE during the 1995-1996 school year. From this population, 200 students were randomly sampled from each group for a total of 400 subjects. The process was as follows: First, both groups were sorted into two segments: those that received free or reduced-priced school lunches (as reported on their test answer documents) and those that did not receive free or reduced-priced school lunches. Of the 386 LEP students, 120 were randomly sampled from those that participated in the free/reduced lunch program and 80 were sampled from those that did not participate in the program. Of the 49,373 regular education students, 120 were sampled from those that participated in the free/reduced lunch program and 80 were sampled from those that did not participate in the program. This strategy was followed in an effort to control the influence of socioeconomic background on comparisons between the two groups. The samples were balanced, 200 in each group, in an effort minimize the impact of differential sample sizes on the results of the confirmatory factor analyses. The sample size of 200 was selected so that a portion of the LEP population (the remaining 186)
could be used as a ‘hold-out’ group for cross-validation purposes. The implications of
this sampling scheme are discussed in the limitation section of this document.

**Definition of Regular Student**

A student is considered a regular student if he/she (a) comes from a home where
the language usually spoken is English, and (b) usually speaks the English language.

**Definition of Limited English Proficient (LEP) Student**

A language minority student is one whose English listening, comprehension,
speaking, reading, or writing proficiency is **below** the average English proficiency level
of English-speaking students of the same age/or grade. This determination is based on
two criteria: (1) the student’s home language and the English language oral proficiency of
the student, and (2) the student’s scores in reading and writing on a English Language
arts achievement test that measures reading and writing proficiency.

The proficiency level of a student is determined by the Language Assessment
Scales (LAS). LAS is a norm-referenced measure which includes oral language and
reading/writing subtests. These tests provide language proficiency classifications of
students from kindergarten through grade 12. In Louisiana, any language minority
student who scores below the proficiency level on the oral, reading, and writing subscales
is classified LEP (LEAP, 1996).

**Analyses**

**Hypothesis 1.** English proficient students will perform better on the
performance-based Written Composition examination than LEP
students.
The confirmation of this hypothesis is expected due to the classifications (regular and LEP) of the subjects included in the study. If students are correctly classified, it is expected that regular students will perform better than their LEP counterparts. However, the literature does not provide clear guidance with respect to whether or not this difference will be uniform across the subscales measured on the Written Composition test. As part of the supplemental analyses of this measure, an attempt will be made to identify which areas are most impacted by English proficiency classification.

To test this hypothesis, a multivariate two sample T-test will be used. This procedure will test for the equality of the vectors of group mean performance on the subskills of the Written Composition test for regular and LEP students (Timm, 1975). This procedure assumes that the data are multivariate normal and that the variance-covariance matrices of the groups are equal. These assumptions were checked empirically. Finally, a 0.05 statistical significance level was used to draw conclusions about the tenability of the hypothesis.

Hypothesis 2. There will be no difference in the performance of English Proficient students and LEP students on the English Language Arts Test.

This hypothesis is based on the authors’ belief that the multiple-choice format of the English Language Arts test will not accurately detect differences in the English proficiency levels of the two groups studied. Should this overall hypothesis be rejected, an attempt will be made to identify subtests on which the groups differ as well as those on which they fail to differ.
As with hypothesis 1, this hypothesis was assessed through a two-sample multivariate T-test procedure. The vector of mean scores on the eight subskills of the test will be compared for the two groups. A 0.05 statistical significance level will be used to draw conclusions about the tenability of the hypothesis.

Hypothesis 3. There will be a significant interaction between assessment approach and English skill level of examinees.

This hypothesis focuses on the three subskills which were common to both the English Language Arts and Written Composition tests: sentence structure, word usage, and writing mechanics. The issue is whether or not the difference in performance across these two tests for these subskills is constant for both groups. That is, is the difference in, say, sentence structure scores on the English Language Arts and the Written Composition tests the same for LEP students as it is for regular students? This hypothesis implies that the distinction between the multiple-choice format and the performance-based assessment format may be more significant for one group than the other. Further, it implies that the distinction may not be constant across subskill areas. This has significant implications for the selection of an assessment format for the two populations studied.

This hypothesis was assessed through a three-step process. First, the subskills measured on both the Written Composition and Language Arts tests (sentence structure, word usage, writing mechanics) were standardized so that each would have a mean of 50 and a standard deviation of 10 for the state’s population. Second, a difference score was constructed for each student for each subskill area:
D1 = PB Sentence Structure Score - MC Sentence Structure Score
D2 = PB Word Usage Score - MC Word Usage Score
D3 = PB Writing Mechanics Score - MC Writing Mechanics Score

Where PB indicates a score based on the performance-based test and MC indicates a score based on the multiple-choice test.

The layout for the two-group analysis is as follows:

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEP Students</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Third, the two group multivariate T-test procedure (Hotelling's $T^2$) was used to test for significant differences in the average D values for the two groups. Also, the possibility of an interaction between group membership and difference was examined.

Hypothesis 4. The factorial structure of the two measures will be invariant for Regular and ESL students.

The previous hypotheses were concerned with differences in mean performance among subtests for the two groups studied. Hypothesis 4 focuses on the relationship among the 13 subtests (8 for the English Language Arts test and 5 for the Written Composition test). Is it the case, for example, that correlations among subtests within assessment type (multiple-choice or performance-based) are stronger than subtest correlations across assessment type? Further, are these relationships constant for the two groups studied? Invariance of the factorial structure of these data would indicate that the
groups are comparable with respect to the constructs measured. However, if the factorial structure varies across groups, this would indicate that the two types of assessments function differently for regular and LEP students. This latter outcome would have implications for issues of bias and accuracy.

The analysis was conducted in several steps. First, Pearson correlations were computed among all subtests for each group studied. Second, exploratory factor analysis was used for each group to identify the number of latent factors underlying correlations among the 13 subtests, and the patterns of loadings on the obtained factors. The initial factor solution was extracted using principal components extraction. The eigenvalue greater than 1.0 rule was used as the basis for determining how many factors to retain. This solution was then rotated, using oblique rotation, to a simplified structure. Subtests were grouped on the basis of their factor loadings.

The third step in these analyses was to test for invariance of the factor structure across the two groups. This was accomplished following a procedure discussed by Byrne (1998): First, utilizing results from the exploratory factor analyses, separate baseline models were established for each group. These are considered to have the same form if the parameter matrices have the same dimensions and the same location of fixed, free and constrained elements. Significant discrepancies could serve as an indication that the model was not invariant across groups. Assuming the baseline models were not dramatically different, the second step was to utilize the multisample capabilities of the LISREL VIII computer program to test for increasing levels of similarity or invariance of the baseline model across the two groups. In the present analyses the strategy was to test for (a) invariant number of latent factors, (b) invariant number of latent factors and
invariant pattern of factor loadings, (c) invariant number of latent factors, invariant
pattern of factor loadings, and invariant latent variances and covariances.

Conclusions regarding the fit or adequacy of the baseline models were based on
practical as well as statistical criteria. Specifically, this involved an examination of (a)
the fitted model for logical consistency with the known structure of the two tests, (b) the
likelihood ratio (LR) chi-square statistic, and (c) the Comparative Fit Index [CFI], and (d)
the Bentler Bonner Index [BBI] (Bentler, 1990). In general, a model was deemed
adequate if (a) it was consistent with substantive predictions, (b) the solution did not
consist of unreasonable or unacceptable parameter estimates (e.g., negative variances or
squared multiple correlations close to zero), (c) the CFI and/or BBI were greater than .90
and/or the LR chi-square statistic, $X^2$, was not statistically significant at the 0.001 level.
Also, while the CFI and BBI help account for the well known sensitivity of $X^2$ to sample
size, we also report $X^2/df$ (where $df$= degrees of freedom)–values greater than 2.0 are
considered to indicate poor fit (For a complete discussion of these issues, see Bollen,
1989).
CHAPTER IV

RESULTS

The results of this study are presented in this chapter in both tabular and narrative form. The hypotheses for this study were tested in the order that they were first presented in Chapter I. All analyses were performed through the use of SAS Analysis System, Release 6.10 (SAS, 1995) and the LISREL VIII computer program (Joreskog & Sorbom, 1993).

Results for Hypothesis 1

Hypothesis 1. Regular students perform better in the performance-based Assessment than limited English proficient students do.

Table 1 presents the means, standard deviations, and effect size statistics for the five dimensions and overall score of the Written Composition Test for the two groups studied. As indicated in this table, regular students obtained an overall written test mean of 11.415, and LEP overall mean of 10.045. The effect size (constructed by dividing the difference in group means by the standard deviation of the LEP group) was 0.36 (values larger than 0.33 are considered to have practical significance; see Borg & Gall, 1989). These suggest that the overall performance of regular students on the Written Composition test was higher than that for LEP students. However, using the 0.33 criteria for determining the meaningfulness of an effect size, it is worth noting that the effect size statistics computed for the individual dimensions do not consistently indicate superior performance for the regular group. The greatest differences appear for Usage and Structure; the smallest differences are for Mechanics and Responsiveness.
Table 1.

Mean Dimension Scores for the English Written Composition Test

<table>
<thead>
<tr>
<th></th>
<th>GROUP</th>
<th></th>
<th></th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regular Group</td>
<td>LEP Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsiveness</td>
<td>M</td>
<td>21.615</td>
<td>20.265</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>5.449</td>
<td>6.769</td>
<td></td>
</tr>
<tr>
<td>Support</td>
<td>M</td>
<td>16.110</td>
<td>14.355</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>4.672</td>
<td>4.968</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>M</td>
<td>6.575</td>
<td>5.075</td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.929</td>
<td>2.510</td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>M</td>
<td>6.385</td>
<td>4.505</td>
<td>.76</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2.095</td>
<td>2.484</td>
<td></td>
</tr>
<tr>
<td>Mechanics</td>
<td>M</td>
<td>6.390</td>
<td>6.025</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2.076</td>
<td>2.243</td>
<td></td>
</tr>
<tr>
<td>Overall Mean</td>
<td>M</td>
<td>11.415</td>
<td>10.045</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>3.244</td>
<td>3.795</td>
<td></td>
</tr>
</tbody>
</table>

The results of a multivariate analysis of variance (MANOVA) for these data indicated a significant main effect of student group ($F (1, 398) = 19.8303; p< .01$), which supported Hypothesis 1. Following the significant main effect of student group, univariate analysis of variance was performed on each of the five dependent variables (Table 2). The results indicate significant differences between regular students and limited English proficient students in 4 of the 5 dimensions. As Table 2 indicates, significant differences were found in dimensions entitled responsiveness to assignment, support/elaboration/organizations, sentence formation, and usage. There was no
significant difference in the dimension of mechanics between regular students and LEP students.

Table 2 also shows $\eta^2$. $\eta^2$ is a measure of the strength of association between a dependent variable and an independent nominal or classification variable. In this study $\eta^2$ is a measure of the degree of association between each of the subtest of the Written Composition test and group membership (regular or LEP). And English proficiency (variable involved in this study (student group). This statistic is also known as the correlation ratio (the between-group sum of squares/the total sum of squares) and can be interpreted as the proportion of variation explained in the dependent variable by the independent classification variable. It has a range of 0.0 to 1.0.

In the present study $\eta^2$ ranged from 0.144 for Usage to 0.007 for Mechanics, which as noted earlier, did not yield a statistically significant F value. Using an arbitrary criteria of 10% explained variance as an index of meaningfulness, only Usage and Sentence Formation yield meaningfully different scores for regular and LEP students. In other words, the differences in the performances of regular and LEP students on the Written Composition test appear to be due largely to these two subtest. Grammatical competence as reflected in Mechanics scores, organization of the composition, and responsiveness to the writing prompt, on the other hand, seem only minimally able to distinguish native speakers of the English language from students for who have known difficulties with the language. The next section of this study indicates if these patterns are upheld with the multiple-choice Language Arts test.
Table 2.

Analysis of Variances for Student Group for the Written Composition Test

<table>
<thead>
<tr>
<th></th>
<th>F Statistic</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate Analysis</td>
<td>19.83**</td>
<td>.047</td>
</tr>
<tr>
<td>Univariate Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Responsiveness to Assignment</td>
<td>4.83*</td>
<td>.012</td>
</tr>
<tr>
<td>2. Support/Elaboration/Organizations</td>
<td>13.24*</td>
<td>.032</td>
</tr>
<tr>
<td>3. Sentence Formation</td>
<td>44.89*</td>
<td>.101</td>
</tr>
<tr>
<td>4. Usage</td>
<td>66.94*</td>
<td>.144</td>
</tr>
<tr>
<td>5. Mechanics</td>
<td>2.95</td>
<td>.007</td>
</tr>
</tbody>
</table>

* p<.05    ** p<.01

Results for Hypothesis 2

Hypothesis 2. There is no difference in multiple-choice scores between regular students and limited English proficient students.

As presented in Table 3, there was a difference in the overall means of regular students and LEP students on the English Language Arts Test. The subskill means were consistently higher for regular students, as compared to limited English proficient students. Regular students had an overall mean of 5.435, whereas limited English proficient students had an overall mean of 4.623. The effect size of 0.48 for the overall mean indicates that the performance of the regular was meaningfully higher than that of
the LEP group. When considering subtest performance, the average for regular students was consistently larger than that of LEP students.

Table 3.

Mean Subskill Scores for the English Language Arts Test

<table>
<thead>
<tr>
<th>GROUP</th>
<th>Regular Group</th>
<th>LEP Group</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vocabulary</td>
<td>M 6.010</td>
<td>5.395</td>
<td>.37</td>
</tr>
<tr>
<td></td>
<td>SD 1.356</td>
<td>1.680</td>
<td></td>
</tr>
<tr>
<td>2. Details</td>
<td>M 6.285</td>
<td>5.300</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>SD 1.242</td>
<td>1.681</td>
<td></td>
</tr>
<tr>
<td>3. Main Idea</td>
<td>M 4.725</td>
<td>3.940</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td>SD 1.719</td>
<td>1.640</td>
<td></td>
</tr>
<tr>
<td>4. Critical Reading</td>
<td>M 5.700</td>
<td>4.705</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>SD 1.774</td>
<td>1.889</td>
<td></td>
</tr>
<tr>
<td>5. Writing Mechanics</td>
<td>M 4.985</td>
<td>4.225</td>
<td>.44</td>
</tr>
<tr>
<td></td>
<td>SD 1.580</td>
<td>1.726</td>
<td></td>
</tr>
<tr>
<td>6. Word Usage</td>
<td>M 4.485</td>
<td>3.375</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>SD 1.867</td>
<td>1.806</td>
<td></td>
</tr>
<tr>
<td>7. Sentence Structure</td>
<td>M 5.435</td>
<td>4.760</td>
<td>.43</td>
</tr>
<tr>
<td></td>
<td>SD 1.522</td>
<td>1.569</td>
<td></td>
</tr>
<tr>
<td>8. Study Skills</td>
<td>M 5.855</td>
<td>5.280</td>
<td>.40</td>
</tr>
<tr>
<td></td>
<td>SD 1.229</td>
<td>1.514</td>
<td></td>
</tr>
<tr>
<td>9. Overall Mean</td>
<td>M 5.435</td>
<td>4.623</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td>SD 1.536</td>
<td>1.688</td>
<td></td>
</tr>
</tbody>
</table>
The results of a multivariate analysis of variance (MANOVA) of the English Language Arts data indicated a significant main effect of student group ($F(1, 398) = 46.1763; p < .0001$). This outcome did not support Hypothesis 2. Following the significant main effect of student group, univariate analysis of variance was performed on each of the eight dependent variables (Table 4). The results indicated statistically significant differences between regular students and LEP on all eight subtests. $\eta^2$ values are largest for Comprehension of Details and smallest for Vocabulary. The subtests Writing Mechanics, Word Usage, and Sentence Structure are common to the Written Composition test. In this instance, however, the mechanics subtest yields a significant difference between the regular and LEP groups. Because of the known differences in the English proficiency of these groups, an important question becomes why the Written Composition test was unable to accurately capture the lack of master of mechanics by the LEP group which was reflected in performance on the English Language Arts test. It is possible that this difference may be due to the limited and unstructured sampling of the mechanics domain in the Written Composition test. That this may not be a problem in the other subtest domains has not been noted in previous literature.

**Results for Hypothesis 3**

Hypothesis 3. There is a significant interaction found between assessment approach and English skill.

As noted in Chapter 3, Hypothesis 3 was assessed through a two group (regular Vs the LEP students) multivariate test based on difference scores. Specifically, I formulated a difference score for each student for each of the subtests that were to the Written Composition test and the English Language Arts test:
Table 4.

Analysis of Variances for Student Group for the English Language Arts Test

<table>
<thead>
<tr>
<th></th>
<th>F Statistic</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate Analysis</td>
<td>46.18*</td>
<td>.105</td>
</tr>
<tr>
<td>Univariate Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Vocabulary</td>
<td>16.23*</td>
<td>.039</td>
</tr>
<tr>
<td>2. Comprehension of Details</td>
<td>44.45*</td>
<td>.100</td>
</tr>
<tr>
<td>3. Comprehension of Main Idea</td>
<td>21.84*</td>
<td>.052</td>
</tr>
<tr>
<td>4. Comprehension of Critical Reading</td>
<td>30.32*</td>
<td>.071</td>
</tr>
<tr>
<td>5. Writing Mechanics</td>
<td>21.09*</td>
<td>.050</td>
</tr>
<tr>
<td>6. Word Usage</td>
<td>36.52*</td>
<td>.084</td>
</tr>
<tr>
<td>7. Sentence Structure</td>
<td>19.06*</td>
<td>.046</td>
</tr>
<tr>
<td>8. Study Skills</td>
<td>17.38*</td>
<td>.042</td>
</tr>
</tbody>
</table>

*P< .05

D1=PB Sentence Structure Score - MC Sentence Structure Score
D2=PB Word Usage Score - MC Word Usage Score
D3=PB Mechanics Score - MC Mechanics Score

where PB indicates a score based on the performance-based test and MC indicates a score based on the multiple-choice test. (As noted in Chapter 3, these tests were rescaled in the total population to have a common mean and a common standard deviation)
The results of these contrasts are presented in Table 5. As reported in this table, all D values are positive, indicating higher performance on the Written Composition test than the English Language Arts test. The largest differences occur for mechanics and the smallest differences occur for sentence structure. The differences are consistently greater for regular students than LEP students. And finally, the smallest contrast in the table occurs for sentence structure for the LEP group. An F test for interaction (1 and 398 degrees of freedom) yielded a statistically significant result at the 0.05 probability level \((F=11.0719)\). This result supports Hypothesis 3. It suggests that the assessment outcomes across the multiple-choice and written response examinations are more similar for LEP students than for regular education students. Also, that this similarity is greatest for sentence structure and least for writing mechanics. This result raises questions about the value of using the expensive and inefficient written assessment for the LEP group. It is possible that the proficiency level is better captured with the objective assessment format.

Table 5.

Mean Differences across PBA and MC

<table>
<thead>
<tr>
<th></th>
<th>Sentence Structure D1</th>
<th>Word Usage D2</th>
<th>Mechanics D3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SD 1.929 1.522</td>
<td>2.095 1.867</td>
<td>2.076 1.580</td>
</tr>
<tr>
<td>LEP Students</td>
<td>M diff. 5.075 -4.760=0.315</td>
<td>4.505 - 3.375=1.13</td>
<td>6.025 - 4.225=1.80</td>
</tr>
<tr>
<td></td>
<td>SD 2.510 1.569</td>
<td>2.484 1.806</td>
<td>2.243 1.726</td>
</tr>
</tbody>
</table>

Note. PBA = Performance Based Assessment. MC = Multiple Choice Format.
Results for Hypothesis 4

Hypothesis 4. The factorial structure of the two measures will be invariant for Regular and ESL students.

For regular students, the correlations for the 13 subtests are presented in Table 6. The correlations are all positive and there is a tendency for the correlations within assessment format (performance-based or multiple-choice) to be larger than the correlations between assessment formats. Further, the correlations among the subtests common to both tests are all in the 0.3 and 0.4 range. These results suggest a significant method of assessment effect. That is, student performance is not only a function of ability in a skill domain, but also the method of assessment as used. Also, as noted above, the multiple-choice test yields results more consistent with known characteristics of the students studied than was the case with the written response test.

The correlation matrix presented in Table 6 was subjected to an exploratory factor analysis. A total of 13 eigenvalues (6.27, 2.17, 0.71, 0.64, 0.54, 0.47, 0.46, 0.41, 0.39, 0.30, 0.23, 0.22, and 0.17) were extracted using principal axis factoring. Using the eigenvalue greater than 1.0 rule, two factors were retained. These eigenvalues, 6.27 and 2.17, together accounted for 65% of the standardized variance. The remaining 11 eigenvalues together accounted only for 35% of the standardized variance.

Table 7 presents the communalities and factor loadings for the 13 subtests on the two retained factor following a oblique rotation to a simple structure. The communalities for all variables are high and the pattern of factor loadings are completely consistent with the structure of the two tests. That is, all subtests for the Written
Table 6: Pearson Correlations among 13 subtests for Regular Students

<table>
<thead>
<tr>
<th>Format</th>
<th>Subtest</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>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBA</td>
<td>1. Responsiveness</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Support</td>
<td>.774</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Structure</td>
<td>.785</td>
<td>.740</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Usage</td>
<td>.688</td>
<td>.692</td>
<td>.756</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Mechanics</td>
<td>.697</td>
<td>.672</td>
<td>.752</td>
<td>.722</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td>6. Vocabulary</td>
<td>.299</td>
<td>.279</td>
<td>.376</td>
<td>.361</td>
<td>.372</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7. Details</td>
<td>.335</td>
<td>.348</td>
<td>.370</td>
<td>.321</td>
<td>.347</td>
<td>.607</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8. Main Idea</td>
<td>.268</td>
<td>.262</td>
<td>.263</td>
<td>.371</td>
<td>.272</td>
<td>.523</td>
<td>.468</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9. Critical Reading</td>
<td>.289</td>
<td>.291</td>
<td>.331</td>
<td>.365</td>
<td>.325</td>
<td>.530</td>
<td>.552</td>
<td>.558</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13. Study Skills</td>
<td>.283</td>
<td>.278</td>
<td>.298</td>
<td>.326</td>
<td>.404</td>
<td>.540</td>
<td>.524</td>
<td>.369</td>
<td>.418</td>
<td>.495</td>
<td>.464</td>
<td>.485</td>
<td></td>
</tr>
</tbody>
</table>
Table 7. **Oblique Two-Factor Solutions for the 13 subtest-data for Regular Students**

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Communality</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responsiveness</td>
<td>.809</td>
<td>-.074</td>
<td>.901</td>
</tr>
<tr>
<td>Support</td>
<td>.777</td>
<td>-.020</td>
<td>.850</td>
</tr>
<tr>
<td>Sentence Structure</td>
<td>.834</td>
<td>.018</td>
<td>.880</td>
</tr>
<tr>
<td>Word Usage</td>
<td>.584</td>
<td>.086</td>
<td>.787</td>
</tr>
<tr>
<td>Mechanics</td>
<td>.778</td>
<td>.112</td>
<td>.772</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.630</td>
<td>.749</td>
<td>.129</td>
</tr>
<tr>
<td>Details</td>
<td>.584</td>
<td>.703</td>
<td>.045</td>
</tr>
<tr>
<td>Main Idea</td>
<td>.531</td>
<td>.688</td>
<td>-.017</td>
</tr>
<tr>
<td>Critical Reading</td>
<td>.575</td>
<td>.704</td>
<td>.016</td>
</tr>
<tr>
<td>Mechanics</td>
<td>.547</td>
<td>.662</td>
<td>.061</td>
</tr>
<tr>
<td>Usage</td>
<td>.510</td>
<td>.628</td>
<td>.077</td>
</tr>
<tr>
<td>Structure</td>
<td>.618</td>
<td>.761</td>
<td>-.048</td>
</tr>
<tr>
<td>Study Skills</td>
<td>.507</td>
<td>.642</td>
<td>-.045</td>
</tr>
</tbody>
</table>

n=200

Composition test loaded on one factor and all subtests for the English Language Arts test loaded on the other factor. (These analyses were also performed with principal factor analysis using squared multiple correlations for the prior communality estimates. The results were unchanged.

For LEP students the correlations for the 13 subtests are presented in Table 8.

Again, the correlations are all positive and tend to be higher within assessment type than across assessment type. Also, for those subtests common to both tests, the correlations were all in the 0.3 range.
The correlation matrix in Table 8 was subjected to an exploratory factor analysis using principal axis extraction. Of the 13 eigenvalues extracted, 2 were larger than 1.0 and were retained. These two accounted for 61% percent of the variance. These results were rotated to a simplified structure using oblique rotation. The final communality estimates and the factor loadings are presented in Table 9. As was the case with regular students, the communalities are all high and the pattern of loadings completely consistently with the known structure of the two tests. (Also, these results did not change when squared multiple correlations were used as prior communality estimates).

Baseline Models: As described in Chapter 3, the second step in evaluating Hypothesis 4 was to fit baseline models to the data for regular and LEP students separately. These would be based on the results of the exploratory factor analysis as presented above. These factor analyses were remarkably consistent for both groups and consistent with the known structure of the tests. Specifically, there were two primary factors and the subtests loaded on the factors in a manner consistent with test structure. These specifications (number of factors and pattern of factor loadings) were used to build the baseline models for each group. Also, the residual terms for all indicators were fixed to be uncorrelated; there were no equality constraints on the factor loadings; and the factor covariances were free to vary.

As presented in Table 10, several goodness-of-fit indices are reported. The first of these is the independence Chi-square statistic for the regular ($X^2(78)=1602.827$) and LEP ($X^2(78)=1476.685$) groups for the 'null model' (Bentler and Bonett, 1980). The null model specifies that the subtests are completely unrelated to one another. The null model
Table 8. Pearson Correlations among 13 subtests for LEP Students

<table>
<thead>
<tr>
<th>Format/Subtest</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>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBA 1. Responsiveness</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Support</td>
<td>.829</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Structure</td>
<td>.537</td>
<td>.584</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Mechanics</td>
<td>.638</td>
<td>.678</td>
<td>.738</td>
<td>.664</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC 6. Vocabulary</td>
<td>.328</td>
<td>.320</td>
<td>.446</td>
<td>.364</td>
<td>.385</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Details</td>
<td>.450</td>
<td>.478</td>
<td>.454</td>
<td>.411</td>
<td>.471</td>
<td>.561</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Critical Reading</td>
<td>.332</td>
<td>.416</td>
<td>.438</td>
<td>.418</td>
<td>.420</td>
<td>.496</td>
<td>.591</td>
<td>.617</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Structure</td>
<td>.230</td>
<td>.357</td>
<td>.343</td>
<td>.353</td>
<td>.381</td>
<td>.427</td>
<td>.503</td>
<td>.441</td>
<td>.488</td>
<td>.409</td>
<td>.422</td>
<td>----</td>
<td></td>
</tr>
</tbody>
</table>
Table 9. Oblique Two-Factor Solutions for the 13 subtest-data for LEP Students

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Communality</th>
<th>Factor Loadings</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Oblique Factors</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1. Responsiveness</td>
<td>.737</td>
<td>-.114</td>
<td>.886</td>
<td></td>
</tr>
<tr>
<td>2. Support</td>
<td>.771</td>
<td>-.053</td>
<td>.885</td>
<td></td>
</tr>
<tr>
<td>3. Sentence Structure</td>
<td>.692</td>
<td>.167</td>
<td>.686</td>
<td></td>
</tr>
<tr>
<td>4. Word Usage</td>
<td>.584</td>
<td>.211</td>
<td>.578</td>
<td></td>
</tr>
<tr>
<td>5. Mechanics</td>
<td>.778</td>
<td>.081</td>
<td>.793</td>
<td></td>
</tr>
<tr>
<td>6. Vocabulary</td>
<td>.554</td>
<td>.672</td>
<td>.042</td>
<td></td>
</tr>
<tr>
<td>7. Details</td>
<td>.630</td>
<td>.662</td>
<td>.116</td>
<td></td>
</tr>
<tr>
<td>8. Main Idea</td>
<td>.606</td>
<td>.718</td>
<td>.031</td>
<td></td>
</tr>
<tr>
<td>9. Critical Reading</td>
<td>.612</td>
<td>.708</td>
<td>.065</td>
<td></td>
</tr>
<tr>
<td>10. Mechanics</td>
<td>.524</td>
<td>.661</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>11. Usage</td>
<td>.494</td>
<td>.664</td>
<td>-.105</td>
<td></td>
</tr>
<tr>
<td>12. Structure</td>
<td>.476</td>
<td>.591</td>
<td>.066</td>
<td></td>
</tr>
<tr>
<td>13. Study Skills</td>
<td>.503</td>
<td>.581</td>
<td>.139</td>
<td></td>
</tr>
</tbody>
</table>

n=200

serves as a base against which to compare alternative models for purposes of evaluating the gain in improved fit. Given a covariance structure of even moderate correlations among variables, the expectation is that $X^2$ value for the null model will be extremely high, thereby indicating excessive malfit. This was the case for both groups: $X^2(78)=1602.827$ for regular group and $X^2(78)=1476.685$ for the LEP group. Also, to address the sensitivity of $X^2$ to sample size, we also report $X^2/df$ ($df=$ degrees of freedom). Again, values larger than 2.0 are considered to indicate malfit.
Table 10.

**Summary of Confirmatory Factor Analysis**

<table>
<thead>
<tr>
<th>Competing Models</th>
<th>X²</th>
<th>df</th>
<th>ΔX²</th>
<th>df</th>
<th>X²/df</th>
<th>Bbi</th>
<th>Cfi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Null Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular</td>
<td>1602.827</td>
<td>78</td>
<td></td>
<td></td>
<td>20.549</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lep</td>
<td>1476.685</td>
<td>78</td>
<td></td>
<td></td>
<td>18.932</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic 2-Factor Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular and Lep</td>
<td>106.306</td>
<td>64</td>
<td></td>
<td></td>
<td>1.66</td>
<td>.934</td>
<td>.972</td>
</tr>
<tr>
<td>Lep Regular Students</td>
<td>186.192</td>
<td>64</td>
<td></td>
<td></td>
<td>2.91</td>
<td>.874</td>
<td>.913</td>
</tr>
<tr>
<td>Final Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular Students</td>
<td>85.748</td>
<td>61</td>
<td>3.44</td>
<td>1</td>
<td>1.41</td>
<td>.947</td>
<td>.984</td>
</tr>
<tr>
<td>Final Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lep Students</td>
<td>66.914</td>
<td>59</td>
<td>100.88*</td>
<td>1</td>
<td>1.13</td>
<td>.951</td>
<td>.990</td>
</tr>
</tbody>
</table>

**Note.** In the analyses, each pair of error covariances was set free, one at a time.

The next step in these analysis was to fit the basic two factor model which resulted from the exploratory factor analyses. The result was a dramatic improvement over the null model for both groups. For regular students education X² dropped from 1602 to 106, and for LEP students it dropped from 1476 to 186. Also, most of the BBI and CFI indices were above .90. Nevertheless, the X² were values were statistically significant and suggested the need to further specify the models. This result was anticipated as a consequence of expected covariances among the measurement errors of the subtests.
Focusing largely on the covariances of the measurement errors of the subtests, I sequentially respecified and reestimated the models until the final best fitting model for each group was obtained in Figures 1 and 2. Error/uniquenesses between subtests of the same measuring instrument were free to covary, resulting in three error covariances for regular students and five error covariances for limited English proficient students.

The results in Table 10 demonstrated a satisfactory fit to the data for regular and limited English proficient students ($X^2 / df = 1.41$, $CFI = .984$, $BBI = .947$; $X^2 / df = 1.13$, $CFI = .990$, $BBI = .951$). The difference in $X^2$ ($\Delta X^2$) for competing models (i.e., nested) models is itself $X^2$-distributed with degrees of freedom equal to the difference in degrees of freedom. It indicates whether the estimated model represents a statistically significant improvement in fit. Evaluation of model fit based on the CFI required the comparison of each model with one in which complete independence of all measurements was posted (i.e., a null model).

Test of Invariance In testing for invariance across groups, the procedures were identical to those used in model fitting. That is, a model in which certain parameters were constrained to be equal across groups was compared with a less restrictive model in which these parameters were free to take on any value. For example, the hypothesis of an invariant pattern of factor loadings was tested by constraining parameters to be equal, and then comparing this model (model 2) with model 1, in which only the number of factors was held invariant across the groups. The hypothesis of an invariant pattern of factor loadings was considered tenable because the difference in $X^2$ ($\Delta X^2(9) = 0.154$) was not significant. The results presented in Table 11 indicate the simultaneous 2-factor solution
Figure 1. Baseline Model of Test Format for Regular Students
Figure 2. Baseline Model of Test Format for LEP Students

for each group yielded a reasonable fit to the data ($X^2/df=2.27$, CFI=.976). These results suggest that the data were well described by the multiple-choice test format and the performance-based assessment format.
Table 11.

Simultaneous Tests for the Invariance of Test Format Structure

<table>
<thead>
<tr>
<th>Competing Models</th>
<th>X^2</th>
<th>df</th>
<th>ΔX^2</th>
<th>df</th>
<th>X^2/df</th>
<th>Bbi</th>
<th>Cfi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Null Model</td>
<td>6304.960</td>
<td>156</td>
<td></td>
<td></td>
<td>40.417</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Number of Factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invariant</td>
<td>263.375</td>
<td>116</td>
<td></td>
<td></td>
<td>2.27</td>
<td>.958</td>
<td>.976</td>
</tr>
<tr>
<td>2 Number of Factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and Pattern of Loadings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invariant</td>
<td>263.529</td>
<td>125</td>
<td>0.154</td>
<td>9</td>
<td>2.11</td>
<td>.958</td>
<td>.977</td>
</tr>
<tr>
<td>3 Model 2 with All</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latent Variances and Covariances</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invariant</td>
<td>263.586</td>
<td>128</td>
<td>0.057</td>
<td>3</td>
<td>2.06</td>
<td>.958</td>
<td>.978</td>
</tr>
</tbody>
</table>

p < .05

To test equality constraints related to the factor variances and covariances, I respecified the measurement model with the equality constraints of the pattern of factor loadings and error covariances to include factor variances and covariances (i.e., the MC test format and the performance-based format). It is noted that the equality of these structural parameters was tested while concomitantly maintaining equality of measurement parameters across groups. Results for testing for equality for factor variances and covariances in the test format model are presented in Table 11. Comparison with those of model 2 in Table 11 reveals a slight increment in fit (CFI = .977 vs. CFI = .978); Model 3 indicates the difference in X^2 values is not significant (ΔX^2(3)=.057).
These findings indicate that the CFI = .978 is an excellent fit to the data and the hypothesis of equality of factor variances and covariances in the test format model was considered tenable across groups. These results indicate invariance in the measurement and structural parameters between regular and LEP students.
CHAPTER V
DISCUSSIONS AND CONCLUSIONS

Discussion

This study investigated the relationship between multiple-choice (indirect) and performance-based (direct) assessments of writing skills for a cohort of 10th grade students enrolled in public schools in Louisiana. The two tests considered, The English Language Arts Test and The Written Composition Test, are parts of a battery of tests (The Graduation Exit Examination) which all public school students in Louisiana must pass before they can be awarded a high school diploma. The present study focused on students designated as regular (native speakers of the English language) and limited English proficient (LEP: non-native speakers of the English language). The purpose of the study was to determine if objective (indirect) and performance-based (direct) assessments of writing in a high-stakes testing program yielded similar results overall and/or for specific skill areas; and to determine if these differences were related to the English proficiency levels of examinees. A related purpose was to determine if these measures reflected similar underlying constructs for the two groups of students studied.

To address these issues, a sample of 200 regular and 200 LEP students were selected from student records of the 1995-1996 administration of the Graduation Exit Examination. As a means of controlling for extraneous factors, the proportion of students receiving free or reduced price school lunches was constrained to be equal in the two samples. Four hypotheses were generated and evaluated using multivariate analysis of variance and confirmatory factor analysis. The results were as follows.
Hypothesis 1. English proficient (regular) students will perform better on the Performance-based Written Composition examination than LEP students.

The results from data analysis addressing this hypothesis indicated that regular students achieved higher scores than LEP students on the Written Composition Test. Regular students showed significant differences in support/elaboration/organizations, sentence structure, and usage in comparison with LEP students. However, there was no significant difference in mechanics and responsiveness to assignment. These results indicated that LEP students obtained about the same mean scores as regular students in writing mechanics (punctuation, capitalization, etc.) and in their ability to discern the requirements of the writing prompt. This outcome may reflect the fact that these dimensions of writing are largely a function of recall and recognition.

The results also indicated that LEP students had problems producing writing samples that displayed the coherence and organization of native speakers. In their written responses, these students, more than regular students, failed to produce conventionalized language consistently, that is, language that was genuinely idiomatic.

Hypothesis 2. There will be no difference in the performance of English proficient students and LEP students on the English Language Arts Test.

The results from data analysis addressing Hypothesis 2 indicated significant differences between regular and LEP students on the English Language Arts Test. Significant differences were found in all eight subtests and regular students consistently scored higher than LEP students. These results did not support the hypothesis. However,
an important observation from this portion of the study was that the results for the multiple-choice test were more consistent with known characteristics of the examinees than was the case with the written assessment. One possible explanation could be the poor content sampling of the written assessment.

Hypothesis 3. There will be a significant interaction between assessment approach and the English skill level of examinees.

The results from data analysis addressing Hypothesis 3 indicated the presence of a significant interaction between English skill level and assessment approach. For three subtests common to both tests, the differences in scores based on the multiple-choice and performance-based assessments were larger for regular students and varied by subtest area. These results imply that the choice of assessment approach carries more consequences for regular students than for LEP students. Also, the greatest similarity of results from the two assessments approaches occurred for sentence structure and the least similarity occurred for writing mechanics.

Hypothesis 4. The factorial structure of the two measures will be invariant for regular and ESL students.

The purpose of Hypothesis 4 was to compare the latent structure of the two assessments for the two groups studied. A baseline model was formulated for each group based on the results of exploratory factor analyses. The patterns of factor loadings were remarkably consistent with the structure of the tests (i.e., subtest from the multiple-choice test loaded on one factor and subtests from the performance-based assessment loaded on the second factor). This was taken as evidence of a significant methods variance.
Confirmatory factor analysis, based on a series of increasingly restrictive models, indicated that the two tests were operating in the same manner in both groups.

Conclusions

The purpose of this study was to discern the extent to which performance-based and objective assessments of writing skills could accurately differentiate between students that were proficient with English (regular students) and those with known English deficiencies (LEP students). On the objective multiple-choice test the regular students consistently performed better than the LEP students for all subtests. However, for the performance-based writing test, while these groups differed on higher-order skills such as sentence structure, usage, and organization, they did not differ on the mechanical aspects of composition such as punctuation, responsiveness to prompt, etc. As suggested above, these results are most likely a function of the poor content sampling of the writing test. However, they could also be a function of the tendency of the LEP students to focus their efforts on memorization and mastery of the rules of composition. In either case, these results suggest that the multiple-choice test was more consistent in its differentiation of these groups than was true of the writing test.

A second central conclusion from this study is based on the results of comparisons of student performance across the two tests. This interaction analysis indicated that regular students performed better, relative to LEP students, on the writing test in the skill areas of sentence structure and usage. In contrast, LEP students performed better in writing mechanics on the writing test than they did on the multiple-choice test. While, as noted earlier, the latter results are likely due to the poor content sampling of the writing test, the former could be attributed to (a) a tendency of the
writing test to magnify differences between the two groups, or (b) the inability of the multiple-choice test to accurately capture the true extent of the differences between the groups. While this issue cannot be resolved with the current study, what is obvious is the fact that the multiple-choice test is far more consistently accurate in its differentiation of these groups than is true of the more time consuming and expensive writing test. In this sense, the multiple-choice test could be said to be more valid.

**Implications**

One implication of this study is that all of us who develop, use, or evaluate educational assessments need to expand our repertoire of strategies and consider alternative models and techniques. We are not without resources in this regard. First, the methodological literature on qualitative or interpretive research methods is rich with advice (Bogdan & Biklen, 1992; Denzin, 1989; Erickson, 1986). Second, and more immediately relevant, models specific to portfolios and other types of less standardized performance assessments serving accountability purposes already exist (Belenoff & Dickson, 1991, Berlak, 1992; Goodman, 1989; Johnson, 1992). Third, models of accountability in other professions such as medicine, law, psychology, and social work have been cited as offering useful alternatives (Darling-Hammond, 1989; Erickson, 1986), and some researchers have suggested strategies for integrating these approaches into more comprehensive systems of accountability that also include more standardized measures (Archbald & Newmann, 1988; Berlak, 1992).

Another important implication of this study is that we need to consider yet another expansion in our delimitation of the concept of validity. Beyond considering the social consequences of assessment-based interpretations and actions, we need to consider
the social consequences of the methods by which we warrant those interpretations and actions. We need to expand our conception of validity to include questions about why particular methods of inquiry are privileged and what the effects of that privileging are on the community. The choice of one or more methods of inquiry must not be taken for granted. It reflects, as Cherryhomes (1988) has noted, effects and exercises of power.

In terms of practical implications, the interaction between English skill and assessment approach found in this study raises a potential concern for the validity of assessment outcomes. PBAs are used for a variety of purposes ranging from low-stakes to high-stakes assessments, from classroom evaluations to certification exams. As validity is relative to the purpose and intended uses of the outcomes, the validity implications of this study differ depending on the intended use of the outcomes of a given PBA.

If, as suggested above, the writing test is superior to the multiple-choice at differentiating students on higher-order skills, then their use in high-stakes competitive settings may be warranted. On the other hand, if a variety of dimensions of performance are being assessed, the more time consuming and expensive performance assessments may not be necessary. However, it may also be the case that the magnified differences between the groups observed for the writing test could be an exaggeration of the true difference and largely a function of culture. This possibility raises the issue of bias and raises questions about the value of these assessments, particularly in high-stakes settings. In contrast, the results for the multiple-choice test were consistent and in directions consistent with the known differences among the students studied. This can only be interpreted as support for their continued use in a variety of settings.
In summary, the premise that proposed high-stakes examination systems with heavy, or possibly exclusive, reliance on performance-based assessments will have beneficial effects also underscores the need to emphasize the evaluation of the consequences of the system. Dunbar, Loretz, and Hoover (1993) observed that “the nation stands poised on the brink of yet another wave of test-based reform, and again we appear prepared to undertake it without sufficient quality control” (p. 302). The quality control that they argue for would include investigations that would address both the evidential and the consequential bases for valid interpretation and use of assessment results that Messick (1989) has articulated. It is incumbent upon the measurement research community to make the case that the introduction of any new high-stakes examination system include provisions for paying greater attention to investigations of the intended and unintended consequences of the system than has been typical of previous test-based reform efforts. As Messick (1992) noted, “This evidence should especially address both the anticipated consequences of performance assessment for teaching and learning as well as potential adverse consequences bearing on issues of bias and fairness” (p. 35).

From existing literature, it is not clear whether the assessment of English skill is, or should be, one of the dimensions of PBA. If English is not an intended dimension of PBA, the results of this study raises questions about the fairness of PBAs. The results of this study can be interpreted to mean that the achievements of limited English proficient students as measured through PBA can be expected to be obscured by English skill, providing regular students with an unfair advantage. As such, the validity of PBA outcomes as indicators of achievement, knowledge, or ability becomes suspect. As
fairness and consequences of assessment are critical validity concerns (Linn et al., 1991; Messick, 1989), this problem needs to be addressed prior to the use of PBA when English skill is not an intended dimension of assessment.

The results of this study may or may not be generalized to other content domains since the manifestations of English skill may be task-related. Also, the interaction between English skill and assessment approach found in this study was based on 10th grade students' performance on a specific MC (the English Language Arts Test) and a specific PBA (the English Written Composition Test) designed to measure sentence structure, word usage, and mechanics. As the range of different forms that PBA might take can be quite broad, including essay items, hands-on projects, lab experiments, demonstrations, and authentic tasks, the extent to which results of this study are generalized to the other variations of PBA is not known. Future research is needed to determine the relationship found in this study with other variations of PBA.

Finally, it is important to note some sampling limitations of the study. In an effort to maintain a balanced group for the confirmatory factor analysis, from a total of 49,373 grade 10 regular students and 386 grade 10 limited English proficient students, 200 students were randomly sampled from each group. While the proportion sampled from the LEP population is high, the proportion sampled from the regular population is low. Although it is expected that these results should generalize to the total population, the question of representativeness should be noted when drawing inferences from this study.
REFERENCES


Cannell, J. (1987). Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average. VA: Friends for Education.


Womer, F. B. (1981) State-level testing: Where we have been may not tell us where we are going. In D. Carlson (Ed.), New directions for testing and measurement: Testing in the states: Beyond accountability. San Francisco: Jossey-Bass.

VITA

Zhenghe Zhang was born on October 24, 1954, in Shanghai, China. He received his bachelor of arts degree in English literature in 1983 from Shanghai Normal University, Shanghai, China. He received the degree of Master of Arts in curriculum and instruction in 1990 from Louisiana State University, Baton Rouge, Louisiana. Currently, he is a candidate for the degree of Doctor of Philosophy in the Department of Educational Leadership, Research and Counseling, which will be conferred in May of 2000.

Mr. Zhang has been an educator for eight years. He has served as an assistant professor in Shanghai Teachers College for five years and an instructor in Evening School of Department of Continuing Education in Louisiana State University for three years. He revised a course in classroom assessment for Independent Study in Louisiana State University. For the past three years, he has served as a psychometrician in Louisiana State Department of Education and has served as a research assistant in the Department of Educational Leadership, Research and Counseling.
DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate:          Zhenghe Zhang
Major Field:       Educational Research
Title of Dissertation: Performance Assessment and English Skill

Approved:

[Signature]
Major Professor and Chairman

[Signature]
Dean of the Graduate School

EXAMINING COMMITTEE:

[Signature]
[Signature]
[Signature]

Date of Examination:

October 22, 1998

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.