1988

The Test of written English: a statistical analysis of validity and reliability

Christina E. English
Portland State University

Let us know how access to this document benefits you.
Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds
Part of the Bilingual, Multilingual, and Multicultural Education Commons, and the English Language and Literature Commons

Recommended Citation

10.15760/etd.5642

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.

Title: The Test of Written English: A Statistical Analysis of Validity and Reliability.

APPROVED BY MEMBERS OF THE THESIS COMMITTEE:

Marjorie Terval, Chairperson

Duncan Carter

Thomas Dieterich

Laureen Nussbaum

This study examines the use of the Test of Written English (TWE), the essay portion of the TOEFL, as an indicator of academic readiness at Portland State University.
Many researchers in the fields of language teaching and testing have argued that objective tests do not adequately reflect students' productive skills. If this view is correct, essay tests should yield additional information concerning a student's language proficiency.

The questions addressed in this study are: (1) Is the TWE a reliable measure of writing proficiency? (2) Is the TWE a valid measure of writing proficiency? (3) Is there concurrent validity between the TWE and objective measures of English language proficiency? (4) Does the TWE evidence predictive validity?

Individual TWE ratings were correlated to determine the degree of interrater reliability. TWE scores were correlated with instructor evaluations based on a six-point scale similar to that of the TWE. TOEFL and MTELP scores, and GPA for the first two terms of academic study, were correlated with TWE scores.

Findings included relatively high interrater reliability: moderate correlations between instructor evaluations and TWE scores; and low to moderate correlations between TWE and objective test scores, the latter indicating that the two types of tests tap different skills.

GPA correlations yielded inconsistent results, i.e., very weak or negative correlations with first-term GPA, and very strong positive correlations with second-term GPA. An
analysis of credit points earned revealed that 50% of the most successful students achieved less than 1 4.0 on the TWE; however, none of the poor students achieved a TWE score of 4.0 or above.

It was concluded that the TWE does yield additional information concerning a student's language proficiency and academic readiness. A TWE rating of 4.0 appears to be sufficient for academic coursework. A score below this level should be weighed with other factors, i.e., other test scores, instructor recommendations and previous academic performance.
THE TEST OF WRITTEN ENGLISH: A STATISTICAL ANALYSIS OF
VALIDITY AND RELIABILITY

By
CHRISTINA ELIZABETH ENGLISH

A thesis submitted in partial fulfillment of the
requirements for the degree of

MASTER OF ARTS
in
ENGLISH: TESOL

Portland State University
1988
TO THE OFFICE OF GRADUATE STUDIES:

The members of the Committee approve the thesis of Christina E. English presented May 6, 1988.

Marjorie Terdal, Chairperson

Duncan Carter

Thomas Dieterich

Laureen Nussbaum

APPROVED:

John R. Cooper, Chair, Department of English

Bernard Ross, Vice Provost for Graduate Studies
ACKNOWLEDGEMENTS

I most gratefully acknowledge those who gave support and assistance in the completion of this thesis. My principle advisors, Dr. Marjorie Terdal and Dr. Duncan Carter, carefully read several drafts of this thesis and provided recommendations and encouragement. I also owe thanks to Dr. Thomas Dieterich and Dr. Laureen Nussbaum who agreed to complete my committee for the oral examinations and offered their valuable suggestions. Dr. Naguib Greis provided information concerning the TWE and rating procedures and gave me the opportunity to participate in a TWE rating session. He also requested materials from Computing Services on my behalf. Lisa Miletich made the collection of test scores as efficient as possible. Diane Fox, Shirley Morrell, and Lynn Tullis took the time and trouble to rate their writing students for me. Diane Fox also guided me in the rating session. Carol Franks and John Sparks offered their study results as well as moral support. To all of the above I am deeply indebted.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>INTRODUCTION</td>
</tr>
<tr>
<td></td>
<td>Statement of Purpose</td>
</tr>
<tr>
<td></td>
<td>Clarification of Terms</td>
</tr>
<tr>
<td>II</td>
<td>REVIEW OF THE LITERATURE</td>
</tr>
<tr>
<td></td>
<td>Problems with Language Proficiency Testing</td>
</tr>
<tr>
<td></td>
<td>Direct and Indirect Measures of Writing Assessment</td>
</tr>
<tr>
<td></td>
<td>The Reliability of Holistic Ratings</td>
</tr>
<tr>
<td></td>
<td>Language Proficiency and Academic Success</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
</tr>
<tr>
<td>III</td>
<td>METHODS AND PROCEDURES</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
</tr>
<tr>
<td></td>
<td>Instrumentation</td>
</tr>
<tr>
<td></td>
<td>Test Limitations</td>
</tr>
<tr>
<td></td>
<td>TWE Rating Procedures</td>
</tr>
<tr>
<td></td>
<td>Restrictions on GPA Data</td>
</tr>
<tr>
<td></td>
<td>Analysis of the Data</td>
</tr>
<tr>
<td>Page</td>
<td>Section</td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>52</td>
<td>Statistical Procedures</td>
</tr>
<tr>
<td>54</td>
<td>RESULTS</td>
</tr>
<tr>
<td>71</td>
<td>Summary</td>
</tr>
<tr>
<td>74</td>
<td>CONCLUSION AND IMPLICATIONS</td>
</tr>
<tr>
<td>74</td>
<td>Discussion</td>
</tr>
<tr>
<td>78</td>
<td>Conclusion</td>
</tr>
<tr>
<td>82</td>
<td>Implications</td>
</tr>
<tr>
<td>85</td>
<td>REFERENCES</td>
</tr>
<tr>
<td>91</td>
<td>APPENDICES</td>
</tr>
</tbody>
</table>
LIST OF TABLES

TABLE                        PAGE

I  Correlation between Instructor Rankings and TWE scores ... 57
II  Paired T-Test for TWE, Instructor Rankings, and TWE Scores ... 60
III TOEFL/TWE Correlations ... 62
IV TWE/TOEFL Section 2 Correlations ... 63
V  TOEFL Means and Standard Deviations ... 64
VI TOEFL Means and Standard Deviations ... 64
VII TOEFL Section 2 Means and Standard Deviations ... 64
VIII MTELP/TWE Correlations ... 65
IX  MTELP/GPA Correlations ... 65
X  TWE/GPA Correlations ... 67
XI  TOEFL/GPA Correlations ... 68
XII  TOEFL/GPA Correlations ... 68
XIII TWE Scores and Credit Points Earned ... 70
<table>
<thead>
<tr>
<th>FIGURE</th>
<th>DESCRIPTION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Scatterplot of the .75 correlation between 2 groups of raters</td>
<td>55</td>
</tr>
<tr>
<td>2.</td>
<td>Scatterplot of TWE scores and instructor rankings for Fall, 1987</td>
<td>58</td>
</tr>
<tr>
<td>3.</td>
<td>Scatterplot of TWE scores and instructor rankings for Winter, 1988</td>
<td>59</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

The Test of English as a Foreign Language (TOEFL) is a widely used and useful instrument for determining an international student's ability to handle instruction in English. One important component of English proficiency is the ability to write clearly.

Until July 1986, the TOEFL measured writing skills through a multiple-choice format which required error recognition, rather than prose generation. While this test format permitted efficient, cost-effective scoring, many would argue that a valid measure of writing needs to include writing (White 1986).

The Educational Testing Service (ETS) has added a 30-minute essay test, the Test of Written English (TWE), to the TOEFL battery. The TWE is a holistically scored test which uses a 1-6 rating scale (see Appendix A).

ETS is continuing studies on the reliability and validity of the TWE and still considers it to be an experimental test. A number of colleges and universities are aiding ETS in their research by administering an experimental version of the essay test, scoring essays in-house, and sending them on to ETS for analysis. Topics
selected for use in the experiment are still being pilot-tested for cultural, and other, biases. Portland State University is a participant in this experiment. Upper-intermediate and advanced ESL students (Levels 3 and 4) at Portland State University have taken the TWE since Winter term, 1987. The Center for English as a Second Language at PSU has recommended that undergraduate students receive a rating of 4 and graduate students receive a rating of 5 on the TWE before pursuing regular coursework.

STATEMENT OF PURPOSE

The purpose of this thesis is to examine the use of the TWE as a measure of academic readiness at Portland State University.

PSU currently uses test scores from the TOEFL and/or the Michigan Test of English Language Proficiency (MTELP), two standardized English proficiency tests, to assess a student's ability to undertake coursework in English. The university recommends scores of 525 (550 for graduate students) on the TOEFL and 80 on the MTELP. These tests include grammar/writing sections, but do not require the student to produce a writing sample.

It has been argued that writing samples are more valid than multiple-choice tests in determining writing proficiency, in part, since writing samples give a much clearer picture of a student's organizational strategies
(White 1986; Benton and Kiewra 1986; Odell 1981; Breland and Gaynor 1979; Cooper 1977). But there are problems with writing assessment based on first-draft writing to a set topic. The student is given no opportunity to revise; and allowances must also be made for variations in writing fluency as a result of a number of factors, including the student's health and state of mind on the testing day, as well as testing conditions. Test design, i.e. topic selection, is an added obstacle to accurate evaluation of student abilities.

Also inherent in the task of language proficiency assessment is the problem of separating language proficiency, general cognitive proficiency, and the level of acculturation (Canale and Swain 1979), all of which interact in the performance of any language task.

Problems also emerge in scoring the essay test. Procedures have been developed to enhance reader reliability for the TWE (Stanfield and Webster 1986); however, ensuring interrater reliability requires carefully controlled conditions (White 1986), which may be difficult to produce, considering the diverse demands made upon ESL instructors.

This study will address the following research questions, to determine whether the Test of Written English yields additional information regarding a student's English competence:
1. Is the TWE a reliable measure of writing proficiency?
   a. What controls are used to ensure intrarater and interrater reliabilities?
   b. Are those controls effective?
2. Is the TWE a valid measure of writing proficiency, that is, does it assess what it aims to assess?
   a. Is there concurrent validity with instructors' evaluations of writing competence?
   b. Is the TWE sufficiently sensitive to measure term-to-term gain?
3. Is there evidence of concurrent validity with objective measures of English language proficiency?
   a. How do scores correlate with overall test scores on the TOEFL and Michigan test?
   b. How do scores on the TWE correlate with scores on the writing/grammar portion of the TOEFL (section 2)?
4. Is there sufficient evidence to support the hypothesis that academic performance can be predicted from test performance?
   a. Is there a significant relationship between TWE scores and GPA in regular coursework at PSU?
   b. Do scores on the TOEFL and/or MTELP yield a strong positive correlation with GPA?

If the TWE is a valid measure of writing proficiency, it should correlate highly with instructors' evaluations of student writing ability, based on in-class work and writing assignments. If writing instructors indicate improvement in student writing, the TWE should reflect this improvement. The reliability of teacher
assessments of writing competence should also be questioned, however. Instructors are exposed to a larger and more representative sample of student writing, but subjectivity may also be a greater factor in their judgments than in the scoring of anonymous essays.

According to Oller (1978), integrative tests should correlate highly with multiple-choice tests if the latter are indeed valid as measures of language proficiency. If correlations among all three tests, the TOEFL, MTELP, and TWE, are high, it would affirm their concurrent validity. However, if TWE scores correlate highly with standardized tests, they may not yield additional information regarding English proficiency. If the TOEFL and MTELP are recognized as valid measures of language proficiency, and if correlations between these measures and the TWE are low, this fact might indicate a lack of validity on the part of the TWE.

Carlson and Camp (1985), reporting on studies based on the essays of 638 college applicants, found strong, but imperfect, correlations between holistically scored essays and TOEFL total scores. They assert that this degree of correlation is evidence that the writing sample contributes additional information, i.e. a demonstration "of cognitive abilities far beyond the mastery of mechanics" (16).
CLARIFICATION OF TERMS

Language Proficiency

The question of what constitutes language proficiency continues to be one of the many hotly debated issues in second language pedagogy and testing.

Bruner (1975) hypothesized a tripartite framework of language competence: a "species minimum" of linguistic competence, communicative competence, and analytic competence. Linguistic competence implies mastery of basic syntactic structures and semantic categories. Communicative competence is defined as the ability to make utterances appropriate to context. Analytic competence involves the successful use of language in the prolonged operation of thought processes based exclusively on linguistic representations.

Canale (1983) also posited three dimensions of language proficiency: basic language proficiency, communicative language proficiency, and autonomous language proficiency. Of course these categories overlap: all include grammatical competence, and communication does not by definition take place in a vacuum devoid of any social or cultural context. However, each dimension highlights a certain aspect of language use. Basic language proficiency in the Canale framework is concerned with the biological
universals required for language development. These universals interact with general cognitive abilities to determine language forms and uses. Communicative language proficiency emphasizes social and cultural context, while autonomous language proficiency focuses on more intrapersonal uses of language, such as problem-solving and creative writing.

Canale (1983) contends that this framework has two important implications for language testing. First, one must not confuse linguistic demands with other cognitive demands made in the assessment of language proficiency. Secondly, test developers must be aware of the contributions of socialization and acculturation to performance on test tasks, including individuals' attitudes toward, and acceptance of, certain tasks.

Oller (1978) and Oller and Perkins (1980) make a case for a single global linguistic ability undifferentiated in component skills. Oller calls this an "expectancy grammar," a mechanism which enables an individual to generate unique models to fit particular situations. Oller's position that language proficiency cannot be meaningfully subdivided into component skills implies that integrative tests of language proficiency are more appropriate than discrete-point tests (Oller 1978; Cummins 1983). Oller includes essay writing as a valid integrative test task.
Integrative/Discrete-Point Tests

For some learners of English there may be a gap between the ability to use English in natural interaction and the ability to perform well on formal test tasks which draw on the application of memorized rules. The first case reflects a stage in language acquisition, while the second is a result of language learning. (Dieterich and Freeman 1979; Krashen 1982). In examining language proficiency tests, a distinction can be made between a test which requires natural use of the language and one which emphasizes adherence to formal rules.

English proficiency tests usually fall into one of two categories: integrative and discrete-point. Integrative tests require the student to produce connected discourse in some meaningful context. The TWE is an integrative test, because it requires students to express themselves in a piece of connected discourse.

Discrete-point tests evoke a structured response of some type. Written tests are frequently constructed in a multiple-choice format. Discrete-point tests allow the test giver to assert control over the grammatical structures to be included. While these tests are easy and cost-effective to score, it is difficult to elicit natural responses. Discrete-point tests focus on learning not language acquisition (Oller 1978; Dieterich and Freeman 1979.) The TOEFL and Michigan Test are discrete-point tests.
Discrete-point tests are often referred to in the literature as indirect measurements of language proficiency, whereas integrative tests are commonly labeled direct measurements. Lloyd-Jones (1987) objects to this labeling, arguing that all tests are to some extent indirect. In testing writing, even a test essay is indirect because it is always a limited sample taken under abnormal conditions. Lloyd-Jones concedes, though, that it is still useful and desirable to distinguish between tests which are based on a writing sample and tests which use an "objective" format to focus on a single language element.

Writing Competence

A full definition of writing competence is beyond the scope of this paper; however, it seems appropriate to attempt some working definition for the purpose of language assessment.

Odell (1981) suggests that writing competence be defined as "the ability to discover what one wishes to say, and to convey one's message through language, syntax, and content that are appropriate for one's audience and purpose" (103).

The TWE scoring guidelines list some of the features of an essay which demonstrates writing competence. The paper must be well organized and well-developed, effectively addressing the writing task, and using appropriate details to support the thesis. It shows unity, coherence, and
progression, and demonstrates consistent facility with the language, which includes syntactic variety and appropriate word choice. In other words, it must illustrate competence on both the syntactic and rhetorical levels (see Appendix A).

**Holistic Rating**

Cooper (1977) defines holistic evaluation as a guided procedure for sorting and ranking written compositions. Raters respond to a written sample by placing the paper within the range of papers produced in response to a given assignment and assigning a rank according to this range.

Charney (1984) defines holistic rating as "a quick impressionistic qualitative procedure for sorting or ranking samples of writing. . . It is a set of procedures for assigning a value to a writing sample according to previously established criteria" (67).

Holistic scoring, unlike analytic or primary trait scoring methods, judges overall writing effectiveness with a single score, rather than assigning separate scores for individual features, such as vocabulary, grammar, content and cohesion.

A holistic rating session usually involves reading a number of sample essays to give the readers the opportunity to internalize the features of each scoring category. In addition, scoring guidelines which make these features
explicit are often at hand for raters to refer to during the rating process.
CHAPTER II

REVIEW OF THE LITERATURE

While language proficiency is but one of the many factors influencing scholastic performance, it is generally believed that the more proficient student has a better chance of performing well in college coursework than one whose English language skills are weak. How effective are our present methods of assessing academic readiness? To what extent can we predict the academic success of ESL students after they exit ESL classes?

This review of the literature will examine some of the questions raised in the testing of writing, including test validity and reliability, rater reliability, and the predictive validity of language proficiency tests.

PROBLEMS WITH LANGUAGE PROFICIENCY TESTING

The most difficult problem with any test of language proficiency is restricting testing to language skills, while limiting the degree of intelligence testing. Cultural differences and differences in individual communication styles are additional confounding factors in measurements of language proficiency.
Intelligence Testing?

White (1986) states:

There is always the possibility, especially strong in tests that seek to measure verbal fluency, of the correlation reflecting a test's ability to measure some third, undefined trait, such as 'general intelligence,' or 'socioeconomic status,' or 'urban sophistication' (184).

The TOEFL was designed as a language proficiency test, and a study conducted by Angoff and Sharon (1971), which compared native-speaker and foreign-speaker performance on the test, seemed to validate this designation. They administered the TOEFL to a group of 71 college freshmen, all native speakers of English. Their test scores were considerably above the mean of 34,774 foreign applications to U.S. colleges. Generally, distributions indicated that the TOEFL was much too easy for the native speakers and inadequate for differentiating among them. Angoff and Sharon also found relatively low correlations (.64) between TOEFL and ACT English scores. This finding was interpreted as additional support for the TOEFL as a measure primarily of language proficiency.

The TOEFL has been modified from a five-part format to a three-part design since the time of the Angoff and Sharon study, and the Educational Testing Service conducted further research in the area of native speaker TOEFL performance. Paul Angelis (1977) reported that a more recent ETS study contradicted earlier findings. The
subjects of this study were 88 college-bound high school seniors. While the total scores achieved by these students indicated that the native speakers performed well on the whole, and had no difficulty with the listening comprehension portion of the test, problems did emerge in both the structure and reading portions. Twenty-two of the 80 test items on the two forms of the test had pass rates below 80% among native speakers, while pass rates for some were as low as 35%. On the reading and vocabulary section of the test, 22 out of the 120 items on the two forms of the test had pass rates below 80% for the native speakers. Students had particular difficulty summarizing and interpreting passages. Angelis claims that the test items requiring these operations make use of mental processes which focus on inference and evaluation. While these processes may indeed be part of what researchers view as "reading comprehension," Angelis contends that language tests can be constructed without including such items.

It appears from these findings that in this instance either cognitive ability, or a standard of formal accuracy beyond the control of the average native speaker, is being tested. It must be remembered that the TOEFL is used to measure the English proficiency of non-native speakers seeking admission to post-secondary academic institutions; therefore the level of language included must be appropriate for all "educated" native speakers of English.
Interpretation of "educated" is perhaps the source of some of these contradictions (Angelis 1977).

The use of time limits in administration of the TOEFL may also contribute to the influence of intelligence on test performance. Nevertheless, if a test measures intelligence as well as language proficiency, Angelis concludes, that fact may enhance its ability as a gauge of academic readiness.

The difficulty of distinguishing between language proficiency and cognitive abilities is but one of the elements affecting test interpretation. Differences in individual communication styles also affect test performance.

**Individual Differences**

Dieterich and Freeman (1979) argue that we must be careful when interpreting tests which emphasize linguistic creativity as a measure of language ability. Language learners who might be characterized as risk-takers are more likely to use the language to express themselves as fully as possible, making more errors in an effort to communicate, while conservative language learners restrict communication in order to avoid technical errors. Dieterich and Freeman claim that these distinctions do not necessarily reflect language proficiency; conservatives may be proficient, but overly cautious, or conversely, lacking in important skills,
but able to use a limited vocabulary and syntax to advantage.

On the other hand, the proficient risk-taker may be willing to experiment with new grammatical structures, or, although less proficient than the conservative learner, may be much more verbose. If performance is scored by grammatical correctness, the conservative will be favored. If length or imagination are decisive scoring criteria, then the risk-taker will have an advantage. This difference in individual style may affect responses to any test task which requires language production.

Social and Cultural Factors

Another factor cited by Dieterich and Freeman is the role played by cultural differences in the performance of creative language tasks. Students from one cultural or national background may be more reluctant to respond at length for reasons unconnected with language proficiency or the linguistic task at hand, whereas others may be much more linguistically assertive in the same situation.

Canale and Swain (1979) also mention the important role played by social and cultural factors:

It is important to keep in mind that language proficiency is only one of many complex and little understood cognitive systems that interact in performance of any language task. Performance on a task is also influenced by socialization and acculturation; the testing situation is a social event (130).
Lloyd-Jones (1987) states that tests which are based on correct standard usage probably measure social identification. He says, "The usage test will probably correlate quite well with what is understood as writing skill even though it is measuring something else" (163).

DIRECT AND INDIRECT MEASURES OF WRITING ASSESSMENT

The Test of Written English has been included as an adjunct to the TOEFL rather than a replacement for the multiple-choice writing/grammar portion of the test. It has been recognized that both direct (essay type) and indirect (objective or discrete-point) measures of writing have specific strengths and weaknesses.

Objective measures of writing proficiency evidence greater reliability (Hirsch 1977; Oller 1978; Breland and Gaynor 1979), but actual writing samples have more face validity (Cooper 1977; Breland and Gaynor 1979; Odell 1981; Benton and Kiewra 1986; White 1986; Lloyd-Jones 1987). Face validity is the way a test appears to test administrators, educators, and researchers. It cannot, of course, be permitted to replace empirical validation, but it does give us some indication of recognized relevance (Harris 1969).

Perkins (1983) states that essay tests of writing have the highest construct validity for the evaluation of overall writing ability, because raters are given the
opportunity to examine a whole test in which the student states, develops and supports a thesis. Construct validity has been defined by Davidson, Hudson, and Lynch (1985) as "measurement of fit to a priori theoretical characteristics" (147). Usually it involves: "(1) a clear statement of theory, (2) an a priori prediction of how the test(s) should behave given that theory, and (3) following administration of the test(s), a check of the fit of the test to the theory." Construct validation of writing tests, therefore, hinges on our definition of writing competence.

Ideally, reliability and validity should both be present in the testing process. A reliable measure must be replicable under equivalent conditions. A valid method of assessment must assess what it claims it does. A valid writing assessment should be able to reflect a writer's "true" abilities (Charney 1984). Dieterich and Freeman (1979) point out, however, that a sample of natural discourse will provide information only on the writer's control of those structures or discourse styles which happen to occur in the sample.

No test, of course, is wholly reliable, since student performance will change from test to test for reasons that may have little or nothing to do with the test (White 1986). Although White argues in favor of writing samples, he adds the caveat that "while first-draft writing to a set
topic is closer to the real writing students do than is an editing exercise, it is by no means the thing itself" (243).

Writing instruction today stresses the fact that first drafts are only the start of the writing process. Writing not only reflects thought, but also facilitates it. Murray (1980) describes writing as an evolutionary process with three stages: rehearsing (prewriting, including notemaking), drafting (a central stage, which gives the writer a chance to figure out what he has to say), and revising (during which the writer makes the writing clearer and more graceful). Often, after drafting an essay, the writer finds that elements need to be reordered, that the writing goes off on tangents, or that major points have been omitted. E. B. White, in The Elements of Style, is a staunch supporter of the use of scissors in writing, and admonishes, "It is no sign of weakness or defeat that your manuscript ends up in need of major surgery" (72).

Taylor (1984) sees writing as a discovery procedure which relies heavily on revision to clarify and refine the discovery. The process of writing helps the writer to shape and refine ideas which are not yet fully formed. Zamel (1983) found that, among the advanced ESL students she studied, changes during revision were most often global, i.e. involving major transformations. Sentences were deleted and added, parts of paragraphs were shifted around, and new paragraphs were formed as thoughts were developed
and expanded. Three, four, and five drafts were not uncommon, particularly among more proficient ESL writers. Zamel also found that the intervention of time between drafts allowed students to see their writing more clearly. Often students needed to read the text aloud to determine what changes were necessary.

Lloyd-Jones (1987) raises the question of variations in the quality of first drafts, also mentioned by Carlson and Bridgeman (1986). Some people are more efficient composers, they point out, while others depend on second and third drafts. Any test sample will favor those who extemporize well, but will not tell us who polishes well.

White (1986) emphasizes that it is unfair to give an important grade based on one mode of discourse, for many students who write well in one mode function poorly in others. A good writing test should therefore require two or more kinds of writing. White goes on to say:

We need to be particularly cautious when we generalize about the meaning of research based on test scores obtained from a test which offers students only one topic, representing one content area, one mode of discourse, one reader audience, and one short period of time for response (187).

Lloyd-Jones agrees that we are all better writers on certain topics and under certain circumstances.

Carlson and Bridgeman (1986) assert that one writing sample is, in measurement terms, the equivalent of a one-item test. Consequently, from the standpoint of validity
and reliability, there are restrictions on the information to be drawn from such a limited sample.

Ideally, then, students should be able to choose among two or three topics on which to write. Yet there are problems with this approach, too. Godshalk, Swineford, and Coffman (1966) found that students do not always select topics which display their best writing; in fact, more proficient students tend to select more difficult topics, and are inadvertently penalized. A menu of topics would also make rating equivalents even less valid.

Consistency of performance across topics has been an unknown factor in the essay test. Carlson et al. (1985), in field tests of the TWE, examined scores across topics. Students wrote on a total of four topics of two types: compare/contrast and chart/graph interpretation. The researchers found a correlation of .83 between scores for one topic type, and correlations of .72, .75, .84, and .69 between the compare/contrast topics and the chart/graph topics. The four correlations represent four language groups: Spanish, Chinese, Arabic, and native English speakers.

With all of the limitations of time, topic equivalencies, and writer error on a specific testing day, there is much support for at least two different writing samples drawn on two separate occasions. Cooper (1977) advocates averaging the grades of two samples of writing
obtained from a student at different sessions. Jacobs (1981) concurs with this view, but suggests that when the composition test is given in conjunction with an objective test battery, one writing sample may be sufficient.

Keeping in mind that a first-draft essay completed within a rigid time limit may not yield a complete picture of a student's writing ability, it does give some indication of his ability to compose and use effective organizational and rhetorical strategies (Mullen 1980; Jacobs, et al. 1981; Odell 1981; Bamberg 1982; Benton and Kiewra 1986). It could also be argued that the successful student should be able to produce a good essay in any of various discourse modes and within certain time constraints (Angelis 1977).

The Relationship of Direct and Indirect Measurements

Do direct and indirect measurements of writing proficiency yield a significant positive correlation? Evidence seems weighted in favor of a relatively strong correlation between the two, although some researchers disagree with this view.

As noted earlier, Oller (1978) stated that if discrete-point tests provide valid assessments of writing ability, they should correlate highly with more direct measures.

It is difficult to say what degree of correlation is satisfactory to indicate concurrent validity between direct
and indirect assessments. Harris (1969) states that we should expect a correlation in the .70 or .80 range between tests.

Some researchers have found relatively high correlations between objective measures which test mastery of mechanics and direct measures (essay performance) of writing ability. Godshalk, Swineford and Coffman (1966) found correlations of .70 and above between essay scores of 646 high school students and scores on the English Composition test, an objective test.

Stiggins (1982) described five studies conducted over a six-year period at various educational levels, from third grade through college. He noted "consistent and relatively strong correlations" (101) between quantitative and qualitative tests and concluded that the two methods assess at least some of the same performance factors. These correlations were below the threshold described by Harris.

Breland and Gaynor (1979) sampled 2,000 entering freshman from four U.S. colleges. Three 20-minute essay scores were compared with three Tests of Standard English (TSWE) scores for tests taken during the 1975-1976 academic year. The TSWE is a 50-item multiple-choice test with a 30-minute time limit. This test is used to assess a student's control of the conventions of standard written English. The researchers found a correlation of .76 between the sum of the three essay scores and the sum of the three
TSWE and the students' SAT verbal scores, which were included as an additional variable. Breland and Gaynor concluded that direct and indirect measurements tend to tap similar skills; but they objected to the use of objective exams on the grounds that they send the message to students that writing is not important.

Lloyd-Jones (1987) states that high correlations have been found between a large vocabulary and writing ability, and students who do well on usage tests (which, in his opinion, actually measure social identification) will often be judged good writers based on their control of standard English. Lloyd-Jones stresses, however, that correlations between the two methods of assessment are far from perfect, and cautions that correlations are not necessarily demonstrated connections.

Bamberg (1982), in contrast to those cited above, found correlations to be consistently low, ranging from .20 to .30, between a multiple-choice test and two essay tasks administered on the same day to the same group. Scores were taken from a Freshman writing program at the University of Southern California over two terms. The tests were part of a final examination which consisted of a 20-minute personal essay, a 40-minute expository essay, and a 35-item objective examination on usage and sentence structure. Bamberg concluded from the weak relationship that the two kinds of tests measure quite different skills.
White (1986) states that the correlation between direct and indirect measures of writing ability hovers around .50. He concludes that a combination of objective and essay tests is more valid than either type alone.

It should be noted that the researchers cited above are comparing scores for native speakers of English. Perkins (1980) found that the TSWE did not discriminate between three different holistic evaluations (pass, low pass, and fail) of ESL compositions. Perkins suggests that tests designed for native speakers should be used cautiously with L2 populations.

Westbrook, Franks, Martin, Seitz, and Fokine (1986) matched carefully controlled essay ratings from Writing 199A classes (freshman composition) of native and non-native speakers of English with scores on the TOEFL and TSWE in a two-year study at PSU. The researchers found low correlations between TSWE and TOEFL scores, and neither scores correlated highly with either essay scores or success in the course. These findings could be due, in part, to discourse-level errors which are not reflected in objective tests.

Kaplan (1980) describes differences in rhetorical patterns used by other cultures which may account for the difficulty non-native speakers of English might have with our linear pattern of organization. One of the examples given by Kaplan is the generally low tolerance for
digression in English discourse. The rhetorical patterns of some cultures not only tolerate digression, but require it.

Carlson, Bridgeman, Camp, and Wanders (1985), contend that while organizational skills often parallel mechanical skills in the writing of native speakers, ESL students may exhibit a greater disparity between the two.

After mastering the syntax of English, the ESL student must understand the logic of the language in order to successfully apply its organizational strategies. A grammar test which measures correct structural recognition, or even production, does not appear to assess this aspect of writing skill effectively.

Selecting the Essay Task

Jacobs, et al. (1981) consider the following criteria to be crucial in the choice of essay topics: the task should be realistic, appropriate, understandable, personal, feasible, reliable, and fair. The task should be realistic in the sense that it is similar to what is required of the student in the "real" world. Appropriateness is defined by Jacobs as free from cultural, socioeconomic, and sexual biases. To be understandable, the task should be stated as briefly and clearly as possible. The task should be personal for the student, that is, of sufficient interest to motivate the student to communicate ideas. The task must be somewhat limited in scope for the sake of both the writer and the reader. Finally, the task should elicit a
representative sample of the writer's ability, in order to be reliable and fair.

Jacobs, et al. suggest that a 30-minute composition task is adequate for students at all but the most basic levels of proficiency to produce a page or more. Length and/or number of words should be de-emphasized, however, or else students will waste time counting words.

Field testing will help ensure that the test topic meets the above criteria. Lloyd-Jones (1987) states that every task should be field-tested and retested after each revision.

THE RELIABILITY OF HOLISTIC RATINGS

One method of essay test scoring currently popular for its economy of time is holistic rating, in which essays are ranked for their overall quality on a numbered scale. Essays are usually ranked on one of the following scales: 1-4, 1-6, 1-8, and 1-10. The TWE is holistically rated on a scale of 1-6.

Despite the relative convenience of holistic scoring compared to so-called analytic scoring methods, such as primary trait scoring, there are those who have some reservations concerning its use. Bizzell (1987) argues that holistically scored tests enable evaluators to relax standards of competency by redefining "good" writing, counting as "good" writing which does not conform to the
rules of Standard English, but perhaps displays other strengths. She opposes the possible use of holistic scoring to right social injustice, i.e., granting admission to students whose English is not adequate for academic work. But Bizzell concedes that "holistically scored writing samples can be used with confidence for the administrative purposes of most testing in American education. They can be used for the coarse sorting of students into basic writing classes, regular classes, and advanced classes" (583).

Carlson and Camp (1985) support the validity of holistic scoring methods. Reporting on studies conducted by the Educational Testing Service to design and pretest writing assessment instruments, the researchers found that holistic scores, discourse-level scores, and sentence-level scores were highly correlated, leading them to conclude that holistic scores alone were sufficient for the consistent evaluation of student essays.

Perhaps the most controversial aspect of holistic rating is the issue of reliability, both intrarater and Interrater. Hours of careful training are required to ensure that ranking standards are adhered to (Cooper 1977). Readers respond differently to various aspects of writing, and if they are not trained stringently enough, raters may allow an emotional response to interfere with judgments. Benton and Kiewra (1986) warn against context effects on
readings: essays are rated low if preceding essays are strong, or high if preceded by poor quality essays.

In spite of training, readers' judgments are strongly influenced by superficial features such as physical appearance (handwriting), word choice, length of essay, and spelling errors. (Nold and Freedman 1977; Charney 1984). Nielsen and Piche (1981) found that the presence of "mature" vocabulary often resulted in higher holistic ratings. Grobe (1981) found essay length and freedom from spelling errors to exert a stronger influence on teacher ratings than syntactic maturity. This fact is unsettling in light of research which classifies errors according to learners' strategies. Proliference of certain kinds of errors, e.g. word form errors, may, in fact, indicate greater writer competence than the absence of such errors (Selinker 1972; Krashen 1982; Carlson and Camp 1985).

Research on holistic scoring has devoted more attention to intrarater reliability than to interrater reliability. One of the few large studies devoted to the latter was conducted at the University of Michigan from August 1979 to March 1980 (Homburg 1984). Thirty-minute writing samples were obtained from international applicants to U.S. colleges and universities. These essays were holistically read and ranked by trained composition readers on a scale of 1 to 10. Intrarater coefficients, based on the correlation between a set of compositions read twice by
one reader, ranged from .874 to .936 for the 3512 compositions.

When readers are unprepared or inadequately trained, interrater reliabilities are low. Diederich, French, and Carlson (1961) conducted a study in which 300 essays were graded by professionals in six fields, including education. Ninety-four percent of the papers received 7, 8, or 9 of the 10 possible grades. One hundred-and-one essays received every grade from 1 to 9, and no essay received less than five different grades from 53 readers. The median correlation between readers was .31; the English teachers in the group had a median intercorrelation of .41. Diederich et al. state, however, that trained College Board readers achieve an average correlation of .70.

Some researchers have found strong interrater correlations, even among untrained raters. Cooper (1977) cites a study at the University of London Institute of Education in which reliabilities as high as .82 were achieved between teams of three raters who were experienced teachers but who were given no training, only minimal instructions, and no chance to discuss among themselves.

The University of Michigan study cited by Homburg (1984) found interrater reliability coefficients of .72 to .93, with a median of .88. Carlson et al. (1985), in pilot-testing the Test of Written English, found interrater reliabilities to be consistently high for the four topics
tested. Paired readers attained correlations of .80, .83, .84, and .85 on the 2552 essays.

Sparks (1988) conducted a reliability study of holistically scored TWEs. Permitting a one-point difference in ranking, a margin of difference compatible with the TWE guidelines (Stansfield and Webster 1986), a reliability coefficient of .90 was achieved. However, although 27 out of 30 pairs of ratings showed no greater than a one-point difference, only eight pairs of scores were identical. Sparks concluded that on a six-point scale, this type of disagreement substantially compromised reliability.

Charney (1984) maintains that the reliability of holistic ratings can be improved if the following conditions are met: readers must be qualified and from similar backgrounds; readers must be "calibrated," that is, trained to conform to established rating criteria; the rating guide must be explicit and appropriate; and readers must work quickly under supervision.

White (1986) outlines six procedures for scoring which should help to control for intrarater and interrater reliability:

1) Raters should be brought together at the same time and place for a controlled essay reading.

2) There must be a scoring guide which defines traits for papers at different points on the scoring scale.

3) There must be sample papers provided to help readers internalize the scoring scale.
4) There should be a table leader who checks scores for uniformity. He must guard against reader drift. A scored sample should be reintroduced after a long session as a check to see if it is scored identically.

5) There should be multiple independent scoring. There should be more than one reader per essay from two different tables. An excellent reading on a six-point scale will have 5% or fewer discrepancies. An average reading will have 7-10% of its scores more than 1 point apart.

6) Records should be kept of readers' scores in order to track their consistency (24).

White concludes that even employing such controls as these, there is no agreement among statisticians about comparative reliabilities of readings, for there are many different ways to compute reader reliability. In general, though, White found that carefully controlled essay readings do yield reasonably reliable scores.

Carlson and Camp (1985) point out that even perfect interrater reliability does not prove the reliability of an essay test. Other factors must also be considered: variation over time, among topics, and from one sample of students to another.

LANGUAGE PROFICIENCY AND ACADEMIC SUCCESS

As noted earlier, the Educational Testing Service recommends institutional flexibility when analyzing TOEFL scores. The ETS manual points out that English proficiency alone does not determine academic aptitude; while low
ability in English may contribute to failure, high ability does not necessarily ensure success (ETS 1987: 19).

Harvey (1979), in his study of the TOEFL as a predictor of GPA, surveyed 78 international students in their first and second terms of regular coursework at Portland State University. He found low correlations of .178 between TOEFL scores and GPA, but made two qualifying remarks. First, Harvey suggested that subtest scores might yield additional information when there is doubt concerning a student's ability to perform well in content coursework; for any one department a particular subtest score might prove more valid than the total score. Secondly, he remarked that the GPA of undergraduate students appears to be more dependent on English proficiency than that of graduate students. He attributes this fact to the more specialized nature of graduate work; undergraduate coursework may necessitate a broader vocabulary and more varied modes of discourse. Another factor to keep in mind when examining test scores and the GPAs of graduate and undergraduate students, Harvey notes, is the dissimilarity of grading practices for these two groups. He concludes that we must not exaggerate the importance of language ability in a scholastic environment. Motivation, academic aptitude, and acculturation are contributing factors in academic performance.
Sharon (1971), in a study sponsored by the Educational Testing Service, compared TOEFL scores and GPA for 978 foreign graduate students at 24 American universities. Twenty-nine percent of the GPAs were based on one semester, 21 percent on two semesters, 19 percent on three semesters, and the remaining 31 on four semesters. Sharon found correlations of .26 using TOEFL as the only predictor of GPA. He concludes that low verbal aptitude is not necessarily a deterrent to success in American graduate schools. He qualifies this observation, however, by stating that it is likely that allowances are made in the grading of foreign students, allowances which contaminate the data and skew the correlation between test scores and academic performance.

Grade point average is, of course, not the ideal measurement of academic success. Cortes (1967) mentions some of the problems with using grades as a reliable indicator of academic success, mainly, that they contain numerous sources of variation. First of all, students take different courses which vary in difficulty. Another important factor is that teachers have different methods of conducting examinations, as well as different criteria for assigning grades. Additionally, some teachers are more lenient in their grading than others. Courseload is another variable of great importance in analyzing GPA (Cortes 1967; Graham 1987); moreover, some students work while attending
school or have time-consuming family or host-family obligations.

When we consider English proficiency as an important determiner of academic success, we are looking at only a part of a student's academic potential. Cortes suggests that potential, which includes intelligence, academic aptitude, and some degree of English competence, combines with motivation to give a clearer picture of what to expect of student performance. Even though motivation is not easily measured, it should be taken into account when a student performs unexpectedly well or poorly.

While the connection between TOEFL score and GPA is tenuous at best, there is some evidence that there is a correlation between TOEFL score and graduate credits earned. Light, Xu, and Mossop (1987) note that if we define academic success by credit hours earned, rather than GPA, we may find a stronger link between TOEFL and academic performance. Graham (1987), reviewing the study by Light, Xu, and Mossop, points out that the .19 correlation between GPA and credits earned is statistically significant but weak.

Graham observes that only students who have demonstrated a high level of English competence are admitted to academic study; if students with lower levels of English proficiency were admitted into academic programs, it is possible that correlations between test scores and academic GPA would be stronger. This hypothesis receives some
support from the research of Gue and Holdaway (1973) whose students had obtained relatively low mean TOEFL scores: 424.6 and 447.8. The correlations between TOEFL score and GPA in this study were relatively high: r = .49 and r = .59. In contrast, Light, Xu, and Mossop (1987) obtained a correlation of only .14 between TOEFL and GPA, but their sample had a high mean TOEFL score: 561.

Graham suggests that proficiency scores above a certain minimum level should not be given too much weight, but the lower the English proficiency is, the greater a factor it is in a student's success. This minimal level probably varies from institution to institution and program to program, in Graham's view. She suggests that each institution conduct its own studies to establish appropriate levels. Graham also recommends that other factors, such as SAT and GRE scores, and high school records, be taken into account in making admissions decisions for foreign applicants. Language test scores should not play a disproportionate role in these decisions, she insists; thus she opposes rigid cutoff scores.

SUMMARY

There are numerous problems inherent in language proficiency testing. Chief among these is the difficulty of separating language skills, intelligence, socio-cultural factors, and individual communication styles (Angelis 1977;
A second major point of debate is the validity and reliability of language proficiency tests. While greater reliability may be achievable with discrete-point tests (Hirsch 1977, Breland and Gaynor 1979), writing samples are intuitively more valid (Cooper 1977; Oller 1978; Breland and Gaynor 1979; Odell 1981; Benton and Kiewra 1986; White 1986; Lloyd-Jones 1987). However, generalizing from a first-draft essay runs counter to writing instruction which places emphasis on the writing process, including re-writing (Murray 1980; Zamel 1983; Taylor 1984; White 1986; Lloyd-Jones 1987).

Essay scoring must be reliable if writing samples are to be accepted as valid proficiency tests. Problems with intrarater and interrater reliability are the strongest argument against the use of writing samples. (Nold and Freedman 1977; Nielsen and Piche 1981; Grobe 1981; Charney 1984;) There are, however, controls which aid in increasing reader reliability (Charney 1984; White 1986).

Predicting GPA on the basis of test scores is complicated by the many factors involved in academic performance: academic aptitude, motivation, course of study, course load, and differences in grading procedures, among others (Cortes 1967; Graham 1987); however, there is probably a certain minimum proficiency level below which
individuals are less likely to succeed as students (Graham 1987).

When making admissions decisions for foreign students it is safest to consider several variables: high school record, SAT and GRE scores, and English proficiency test scores. Some researchers believe that language test scores should not be overemphasized, and rigid cutoff scores should be eliminated (Harvey 1979; Graham 1987).
CHAPTER III

METHODS AND PROCEDURES

In this chapter the process of data collection and data analysis will be described. Data collected for this study includes TWE, TOEFL, and MTELP scores for students enrolled during the 1987 Winter, Spring, and Fall Terms, and during Winter, 1988; TWE ratings given by each of the raters involved in scoring the essays; writing instructor evaluations for students enrolled in Fall, 1987, and Winter, 1988, Level 3 and 4 writing classes; and first- and second-term GPA for students meeting the GPA study criteria. Following a description of the subjects, background information on the three tests, the TOEFL, MTELP, and TWE, will be provided, including statements on reliability from test sponsors. Limitations on coursework to be included in GPA correlations will also be discussed. A rater training session at PSU will be examined for compliance with ETS-prescribed rater training procedures. Finally, statistical procedures will be clarified.

SUBJECTS

The subjects of this study were 120 full-time students, enrolled in Level 3 and 4 ESL classes at PSU and
continuing studies at Portland State after completion of the ESL program. Some were enrolled only for one or two of the four terms. Unfortunately, relatively few of the subjects were enrolled in graded academic coursework in succeeding terms; consequently, grade point averages were available for only 21 of the 120 students in the sample. Only 18 subjects had completed two terms of post-ESL study at PSU; the remainder completed only one term before the conclusion of this study.

Research was limited to students in the upper two levels of ESL study, because students below this level are not required to take the TWE.

Level 3 students are considered to have upper-intermediate ability in English. Generally, they are not proficient enough to undertake regular coursework. They have mastered a basic vocabulary and grammar, but lack a broad enough knowledge of syntax and sufficient vocabulary to perform well in an academic environment. At PSU, students are permitted to enroll in one non-ESL class while in Level 3.

Students in Level 4 are developing competence in English, but generally still make frequent mistakes in usage and syntax. Level 4 is the final term of ESL in the PSU program. Instruction is geared toward academic needs such as writing research papers, note-taking, and basic computer
skills. Students may take two non-ESL courses while enrolled in Level 4.

**INSTRUMENTATION**

**Test of Written English**

The Test of Written English (TWE) is an adjunct to the TOEFL (Test of English as a Foreign Language), developed by the Educational Testing Service in response to numerous requests that a writing sample be incorporated into the TOEFL. The TWE, added in July, 1986, is a thirty-minute writing test in which students are given a set topic, then advised to make notes and organize their essays before beginning to write. Essays are then scored holistically, that is, "for overall effectiveness of the communication, rather than for separate analytical criteria such as structure, spelling, punctuation, and word usage" (Stansfield and Webster 1986: 17). Essays are rank ordered on a scale of 1 to 6, with six as the highest score (see Appendix A for scoring guidelines). Normally, each essay is ranked by two readers. After all essays are rated separately by both raters, scores are compared. If there is a one-point disagreement, the two scores are averaged, e.g. $3 + 4 = 3.5$. If there is a discrepancy greater than one point, the essay is given to an arbitrator. At PSU essays are sometimes discussed and consensus is reached, rather than taking an average of two scores.
Development of the Writing Test

Pilot studies were conducted by the TOEFL research committee to find the best possible essay topics. Seventy topics were submitted; of these, eleven were selected for pretesting. Six of the eleven were adopted, with two accepted after minor revisions. Carlson and Bridgeman (1986) explain the selection process. The topics chosen had to be fair, that is, free from cultural biases. This aspect alone, as they observe, limited the field drastically: controversial social norms, technological advances, or social systems can evoke emotional responses which might influence a student's ability to write coherently. Similarly, the topic should not suggest an orientation which the student is expected to share, i.e. American or middle class. At the same time, the assignment must be sufficiently compelling to stimulate the test-taker to want to communicate ideas. For the readers' sake, also, the topic should provide sufficiently diverse responses to prevent tedium; otherwise scoring accuracy could be affected. However, responses need to be sufficiently uniform to compare on an evaluative scale. The test must not include a reading passage, of course, for then it could be argued that reading skill is also being tested. Finally, the topic should elicit the kind of writing demanded in an academic setting, rather than pure narrative.
Carlson and Bridgeman surveyed faculty members in 190 departments at 34 universities in the United States and Canada to determine which academic writing skills were deemed most important. Although the ability to argue in favor of one point of view in contrast to another was regarded as very important for undergraduates in general, for business administration students, and for psychology majors; argumentation was viewed as less important in engineering, computer science and chemistry, three fields which attract a high proportion of international students.

In the sciences, interpretation of a chart or graph was considered to be among the most important writing tasks. Thus, the two types of topics chosen by the TOEFL committee represent these two discourse modes: argumentation and chart/graph interpretation (see Appendix C for sample topics).

**TOEFL**

The TOEFL is an internationally recognized English proficiency examination. Introduced in 1963 as a 5-part test, it was redesigned in the late 1970s and now consists of three sections: (1) Listening Comprehension, (2) Structure and Written Expression, and (3) Vocabulary and Reading Comprehension. The TOEFL uses a multiple-choice format and is given in a single three-hour session. Intra-test and inter-test correlations are high, .92 to .95, according to the ETS manual and Carlson, et al. (1985).
The total possible score on the TOEFL is 677. Colleges and universities in the United States and Canada often require minimum TOEFL scores for foreign student admission. At PSU, undergraduates must have a score of 525 for admission as a regular full-time student; for graduate students the minimum is 550. Individual departments may set higher minimums; for graduate study in English the required score is 600.

The Educational Testing Service has repeatedly cautioned that TOEFL scores are indicative of English proficiency only, and should not be used to predict academic performance (TOEFL Test Manual). The test developers provide the following guidelines for the use of test scores:

1) supplement TOEFL scores with other relevant information pertaining to academic aptitude, e.g. high school records;
2) consider the standard error of measurement when determining cut off scores; 3) consider subtest scores; 4) consider specific English language requirements of different fields of study; 5) consider the impact of English proficiency on other standardized test scores, e.g. GRE; 6) do not interpret TOEFL scores as measurements of academic aptitude (ETS 1987:17).

The Michigan Test

The Michigan Test of English Language Proficiency (MTELP) is given quarterly to students in PSU's ESL program. It is a 100-item objective test in three parts: grammar,
vocabulary, and reading comprehension. Students are given 1-1/2 hours for the test, which includes 15 minutes of oral instruction. The Michigan Test Manual cites correlations with TOEFL scores ranging from .80 to .94, based on results obtained from four American universities from August 1965 through February 1966.

Since the MTELP has several forms (at least 12 of which are used at PSU), an equation table is provided in the test manual to enhance test score equivalency.

The English Language Institute, sponsor of the MTELP, states: "Scores indicate relative ability to pursue academic study in an English language university" (Michigan Test Manual, 1962). The following are suggested norms:

- 90 and above - full-time courseload
- 85 - 89 - 3/4 courseload
- 80 - 84 - 1/2 courseload
- 70 - 79 - 1/2 - 1/3 courseload
- 69 and below - not proficient enough to undertake academic coursework

Portland State University recommends that a student attain a score of 80 on the MTELP before exiting the ESL program.

TEST LIMITATIONS

TOEFL and MTELP

It is common for a student to take the TOEFL and Michigan Test several times when the student's goal is
acceptance at an American college or university. In order to prevent students from bettering scores by remembering and researching test questions, both the TOEFL and MTELP are available in several forms, i.e. containing different test items. Sponsors of both tests claim high intertest correlations, in the .90s (Michigan Test Manual and TOEFL Manual), but it should be noted that students in the sample took four different forms of both tests, and no exact figures are available on test equivalencies for these particular tests.

**TWE**

As in the case of the TOEFL and MTELP, the question of test equality must be considered for the TWE. The test sponsors seek to avoid student familiarity with test items; consequently, essay topics vary from one test session to the next. Carlson and Bridgeman (1986) cite a .83 correlation between two topics of the same type (see Review of the Literature); however, this relationship might be weakened across four topics.

Furthermore, unlike the TOEFL and MTELP, the scoring process for the TWE involves subjective judgments on the part of the raters. While we have seen that carefully controlled holistic ratings yield reasonably reliable scores (White 1986), reliability is further compromised if the same instructors do not participate in each rating session. At PSU, writing instructors generally score TWE tests, and
precautions are taken that they do not score the tests of their own students, since there could be some bias on their part; however, teaching assignments vary from term to term, and this results in some rotation of TWE raters. Among the tests included in this study, at least eight different instructors rated the 114 TWEs. Interrater reliability was reasonably high, .75 (see Results and Discussion), indicating that paired instructors had little difficulty reaching consensus; however, it is difficult to determine whether or not a different pair of readers would arrive at the same determination.

**TWE RATING PROCEDURES**

I participated in a TWE rating session at PSU during Winter term, 1988. This session conformed roughly to procedures established and employed by the Educational Testing Service. The procedures outlined by ETS include reading of sample essays, followed by discussion of each essay and how similar essays should be treated. After scoring criteria have been internalized, rating of the TWE essays proceeds. Each essay is scored by two raters working independently. The scores of two raters are averaged if there is a one-point discrepancy. When there is a disagreement of more than one point, an outside rater scores the essay.
There were three raters at the PSU rating session. We first received a copy of the TWE rating guidelines. Sample essays representing each of the six categories were then read. Next, ten practice essays were read and scored. These ratings were compared; for each essay most were within a point of the other two ratings. We discussed our reasons for giving discrepant scores. We subsequently rated another ten essays. This time, there was greater consensus among the three raters. These preliminary readings lasted approximately 1 1/2 hours. Due to time constraints we were forced to begin rating the TWEs the following day. At the beginning of the second session we rated and discussed sample essays for approximately 30 minutes before rating the 30 TWE essays. We three raters gave identical scores for fewer than half of the essays; for over half of the essays there was a one-point discrepancy between at least two of the three raters, and two of the essays were given to an outside arbitrator. One of these had a discrepancy of two points, and the other essay did not address the topic and thus was not rated at all, in accordance with TWE guidelines.

The scoring process at PSU differs for practical reasons, from the established procedure in a few specifics. At ETS centers a table leader is assigned who supervises the six to eight readers at each table, randomly reading essays scored by the readers. If a table leader disagrees with a
score, s/he will discuss it with the reader. Sample essays are also periodically introduced to monitor reader accuracy. Due to scheduling conflicts, it is not feasible for all scorers to meet together at PSU; as a result, essays are scored in pairs. The strict controls enforced by the table leader are absent. ETS cites reader reliabilities ranging from .85 to .87 for the 110,000 TWE essays read under optimal conditions. Reliabilities for TWE essays scored by individual institutions are probably somewhat lower.

RESTRICTIONS ON GPA DATA

As mentioned earlier, there are numerous sources of variation which affect GPA, including the number of credits carried per term and the type of courses in which the student is enrolled. Some courses make few demands on a student's English language proficiency, for example, those involving physical or artistic skills and courses in the student's native language. For this reason, physical education courses, music performance courses, studio art classes, and courses in the student's native language were excluded from the calculated GPA. For the purpose of this study, a minimum of 8 credits was required, excluding the aforementioned types of courses. If a student was enrolled in only one course during a particular term, grades earned during the following term were included in the analysis.
Therefore, first and second term GPA represents, in some cases, first and third terms or second and third terms.

ANALYSIS OF THE DATA

All TOEFL, MTELP, and TWE scores for tests administered at Portland State University are kept on file in the Center for English as a Second Language office. Scores for the three tests were obtained for students enrolled in levels 3 and 4 from Winter, 1987 through Winter, 1988, with the exception of Summer term, 1987, during which no TWE was given. The TWE given Winter, 1987 was the first Test of Written English administered at PSU. The TWE administered during Winter, 1988 was not sponsored by ETS, due to a missed deadline, but adhered to TWE testing and scoring guidelines.

Writing instructors for levels 3 and 4 in Fall, 1987 and Winter, 1988, were asked to rate their students' overall writing proficiency, as demonstrated by in-class and take-home assignments, on a six-point scale tied to the TWE scale (see Appendix B) without referring to TWE scores. During Winter, 1988 teachers rated students before they took the TWE. This was not the case during Fall term, but instructors were cautioned to avoid consulting TWE scores when determining ratings, a precaution which should have minimized contamination of the data.
Although categories on the TWE scale were reordered and reworded, criteria defining each category were not altered significantly in creation of the instructors' rating scale. One instructor commented that she had difficulty assigning rankings because she found the collection of criteria for each category so arbitrary. Her comments were that some students organize and develop well, but don't support with evidence, and she considers grammar and word choice individual problems unrelated to other criteria. This instructor pointed out that she had experienced the same difficulty in rating TWE essays.

TWE essays are scored at PSU, usually by writing instructors. Ratings given by individual instructors are kept on file to track scorers' consistency. Eighty-five individual ratings were obtained and correlated to determine the degree of interrater reliability at PSU. These scores were taken from the first three TWE tests.

Computing Services provided a computer run of all ESL students enrolled during the period under study. Grade point averages were extracted from the computer run for the first two terms of post-ESL study.

All test scores and grade point averages were entered into an IBM 3161 computer. SPSSx software, a statistical program, was employed to derive means, Pearson correlation coefficients, frequencies, standard deviations, and plots for the variables.
The total TOEFL score and the MTELp scores for each term were correlated with the TWE taken during the same term. Section 2 of the TOEFL, the writing and grammar portion of the test, was also correlated with the TWE and with the total TOEFL score.

A paired t-test was used to determine if term-to-term improvement was reflected in TWE scores. Means for Fall 1987 TWE scores, TOEFL scores, and instructor evaluations were compared with means for Winter, 1988.

TWE Scores for Winter, Spring, and Fall, 1987 were correlated with GPA in coursework taken in the first two terms following ESL study. MTELp and TOEFL scores were also correlated with derived GPA. Grade point averages were multiplied by the number of credits earned to determine credit points per student.

For students enrolled in Level 3 and 4 ESL coursework in Fall, 1987 and Winter, 1988, instructors' rankings of student writing proficiency were correlated with TWE ratings.

STATISTICAL PROCEDURES

The Pearson product-moment coefficient, or Pearson r, was used to analyze the data. This correlational technique was designed for use with interval data. The formula for computing correlation coefficients converts scores to the
same type of scale, thus enabling the researcher to compare unlike scores such as TOEFL and TWE scores.

Sharp (1979) gives a rough gauge for interpreting a coefficient of correlation:

- .85 to 1.00: High
- .50 to .84: Moderate
- 0 to .49: Low

From correlation coefficients one can draw conclusions only in regard to strength of a relationship, not about cause and effect. It should be noted that unless a correlation is particularly strong, and based on a large sample, the degree of correlation could be due to sampling error. For this reason, it is useful to set a significance level ($P = \cdot05$). For this study the significance level was set at .05, that is, $P$ must be .05 or less. The significance level allows us to make predictions or generalizations from a small sample.
RESULTS

Interrater/Intrarater Reliability

The first question addressed in this study was the degree of TWE interrater and intrarater reliability at PSU. The interrater reliability portion of this study included 85 pairs of TWE ratings from 4 pairs of raters. These ratings were taken from Spring, 1987, Fall, 1987, and Winter, 1988. Individual ratings were not kept on file for Winter, 1987, the first TWE administration at PSU. Other ratings not included in this portion of the study are those which were determined through discussion and consensus, a method which achieves an interrater reliability of 1. Implicit in this reliability figure is no guarantee of accurate judgment on the part of the raters, merely agreement.

Ratings from the three rating sessions were pooled in order to increase sample size. The four pairs of raters were divided into two groups; Group 1 represents four raters and Group 2 represents their four partners. The two groups achieved a positive Pearson correlation coefficient of .75. Figure 1 gives the plot for the 85 ratings.
Figure 1. Scatterplot of the .75 correlation between 2 groups of raters (85 ratings).
Forty-three of the 85 ratings, or 50.5% were identical. Only 2 of the ratings indicated a disagreement in excess of 1 point, and the balance of 42 ratings, demonstrated a one-point difference.

It should be noted that these are pooled ratings from three rating sessions; reliabilities for each session may be higher or lower. A correlation of .75, though reasonably high, falls short of the .85 to .87 reliabilities claimed by ETS (ETS 1987).

As mentioned by White (1986) and others, it is difficult to ensure intrarater reliability when essays are scored in pairs, without the benefit of an outside arbitrator who monitors scoring consistency. Intrarater reliability probably suffered from the lack of group rating sessions in which a Table Leader reintroduces scored essays and guides against reader drift.

Correlations between Instructor Evaluations and TWE Scores

The second question to be addressed in this study is the validity of the TWE as determined by correlations with teachers' evaluations of writing ability and the accurate measurement of term-to-term gain.

Correlation coefficients for teachers' assessments of writing skill and TWE scores were moderate as indicated in Table II.
TABLE I
CORRELATIONS BETWEEN
INSTRUCTOR RANKINGS AND TWE SCORES

<table>
<thead>
<tr>
<th>Term</th>
<th>Pearson r</th>
<th>Cases</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 1987</td>
<td>.50</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Winter 1988</td>
<td>.39</td>
<td>46</td>
<td>.004</td>
</tr>
</tbody>
</table>

A moderate and statistically significant correlation was found between instructor rankings and TWE scores for both terms.

Figure 2 gives the scatterplot for the .50 correlation for Fall term; Figure 3 gives the plot for the .39 correlation for Winter term.

In order to determine if term-to-term improvement was reflected in TWE scores, and if so, if this improvement was substantiated by an increase in the means of instructor evaluations and TOEFL scores, a paired t-test was used. The results of the t-test are presented in Table II.

All three measures of language proficiency indicated term-to-term improvement; however, for the TOEFL scores, this improvement was not statistically significant. The difference in the means of the two TWE tests was -.69; the difference in means between the instructor rankings was -.86. These figures indicate that whereas there was only moderate agreement between instructor rankings and TWE scores for the two terms, both instruments appear to be
Figure 2. Scatterplot of TWE scores and instructor rankings for Fall, 1987.
Figure 3. Scatterplot of TWE scores and instructor rankings for Winter, 1988.
### TABLE II
**Paired T-test for TWE, Instructor Rankings, and TOEFL**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>CASES</th>
<th>MEAN</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
<th>(DIFFERENCE MEAN)</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
<th>2-TAIL CORR. PROB.</th>
<th>VALUE</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALL TWE</td>
<td>18</td>
<td>3.19</td>
<td>.94</td>
<td>.22</td>
<td>-.69</td>
<td>.94</td>
<td>.22</td>
<td>.460</td>
<td>.055</td>
<td>.006</td>
</tr>
<tr>
<td>WINTER TWE</td>
<td></td>
<td>3.88</td>
<td>.87</td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
<td>-3.13</td>
<td>-3.13</td>
<td>.006</td>
</tr>
<tr>
<td>FALL INSTRUCTOR RANKING</td>
<td>18</td>
<td>3.00</td>
<td>.78</td>
<td>.18</td>
<td>-.86</td>
<td>.50</td>
<td>.12</td>
<td>.834</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>WINTER INSTRUCTOR RANKING</td>
<td></td>
<td>3.86</td>
<td>.92</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td>-7.18</td>
<td>7.18</td>
<td>.000</td>
</tr>
<tr>
<td>FALL TOEFL</td>
<td>12</td>
<td>483.33</td>
<td>25.79</td>
<td>7.44</td>
<td>-15.75</td>
<td>26.83</td>
<td>7.74</td>
<td>.469</td>
<td>.124</td>
<td>.067</td>
</tr>
<tr>
<td>WINTER TOEFL</td>
<td></td>
<td>499.08</td>
<td>26.27</td>
<td>7.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
sufficiently sensitive to measure term-to-term gain.

It should be noted that instructor assessments are also subject to inconsistencies. Although writing instructors have access to a larger body of the students' work than is represented in one first-draft writing sample, instructors are not free from subjectivity which could bias judgments. It is perhaps easier to perceive improvement in writing than it is to rank students on a rating scale, one which may be flawed in its attempts to categorize writing ability.

**TOEFL/TWE Correlations**

The third question this study addresses is the concurrent validity of the TWE. How does it correlate with TOEFL and MTELP scores? The sample included 120 students, for these 159 TWE scores were collected, representing 73 students who took 1 TWE test; 38 students who took 2 TWEs; 2 students who took 3; and 1 student who took 4 tests (he repeated both Level 3 and Level 4). Six students had no TWE scores; they were enrolled in Level 4 only and were absent on the test date.

TOEFL/TWE correlation coefficients and significance levels are given in Table III.
TABLE III
TOEFL/TWE CORRELATIONS

<table>
<thead>
<tr>
<th>Term</th>
<th>Cases</th>
<th>Pearson r</th>
<th>Significance Level (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 1987</td>
<td>29</td>
<td>.46</td>
<td>.013</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>32</td>
<td>.36</td>
<td>.024</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>29</td>
<td>.19</td>
<td>.169</td>
</tr>
<tr>
<td>Winter 1988</td>
<td>28</td>
<td>.42</td>
<td>.017</td>
</tr>
</tbody>
</table>

With a significance level of .05, there appears to be a moderate to low positive correlation between the two tests, indicating that overall language proficiency, and/or the ability to perform well on the TOEFL, is inadequately reflected in TWE scores (Carlson and Camp 1985).

Correlations for Section 2 of the TOEFL with the TOEFL total are expectedly high: .72 for Winter 1987, .85 for Spring, 1987, .68 for Fall, 1987, and .89 for Winter, 1988. These correlations are weakest in Fall, 1987.

Correlations for Section 2 of the TOEFL, the structure and writing section, and the TWE are considerably lower than for the TWE and total TOEFL score and above the .05 significance level in all cases, as demonstrated in Table IV.
### TABLE IV

**TWE/TOEFL SECTION 2 CORRELATIONS**

<table>
<thead>
<tr>
<th>Term</th>
<th>Cases</th>
<th>Pearson r</th>
<th>Significance Level (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 1987</td>
<td>23</td>
<td>.34</td>
<td>.056</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>28</td>
<td>.30</td>
<td>.053</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>26</td>
<td>.06</td>
<td>.376</td>
</tr>
<tr>
<td>Winter 1988</td>
<td>22</td>
<td>.40</td>
<td>.033</td>
</tr>
</tbody>
</table>

This weak correlation appears to indicate that the two tests of writing competence make different demands on the test-taker.

The very low correlations for Fall, 1987 in all three cases, TOEFL/TWE, TOEFL/Section 2, and Section 2/TWE, appear to be anomalous, the cause could lie in testing conditions for either or both of the tests, or the team of TWE raters for that term. Interestingly, TOEFL intertest correlations are also lowest for Fall, 1987, and the Section 2 score is the highest, which throws suspicion on the specific TOEFL test used in the Fall, 1987, test administration.

Because of the low correlations found for Fall, 1987, TOEFL and TWE mean scores were tabulated for the four terms. An examination of Tables V, VI, and VII reveals that total TOEFL scores and Section 2 TOEFL scores were highest during Fall, 1987, but not significantly so.
TABLE V
TWE MEANS AND STANDARD DEVIATIONS

<table>
<thead>
<tr>
<th>Term</th>
<th>Cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 1987</td>
<td>32</td>
<td>3.53</td>
<td>.80</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>39</td>
<td>3.29</td>
<td>.92</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>43</td>
<td>3.31</td>
<td>.87</td>
</tr>
<tr>
<td>Winter 1988</td>
<td>47</td>
<td>3.38</td>
<td>.79</td>
</tr>
</tbody>
</table>

The breakdown of TOEFL scores for the four terms is revealed in Table VI.

TABLE VI
TOEFL MEANS AND STANDARD DEVIATIONS

<table>
<thead>
<tr>
<th>Term</th>
<th>Cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 1987</td>
<td>29</td>
<td>475.4</td>
<td>41.7</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>32</td>
<td>475.6</td>
<td>44.7</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>29</td>
<td>507.7</td>
<td>32.3</td>
</tr>
<tr>
<td>Winter 1988</td>
<td>28</td>
<td>505.4</td>
<td>30.9</td>
</tr>
</tbody>
</table>

Table VII gives scores for Section 2 of the TOEFL.

TABLE VII
TOEFL SECTION 2
MEANS AND STANDARD DEVIATIONS

<table>
<thead>
<tr>
<th>Term</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 1987</td>
<td>46.8</td>
<td>5.3</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>45.3</td>
<td>4.9</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>50.2</td>
<td>4.9</td>
</tr>
<tr>
<td>Winter 1988</td>
<td>48.8</td>
<td>4.8</td>
</tr>
</tbody>
</table>
Correlations between the Michigan Test and the TWE were similar in degree to those found between the TOEFL and TWE (Table III) as demonstrated in Table VIII.

**TABLE VIII**

**MTEL/P/TWE CORRELATIONS**

<table>
<thead>
<tr>
<th>Term</th>
<th>Cases</th>
<th>Pearson r</th>
<th>Significance (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 1987</td>
<td>25</td>
<td>.48</td>
<td>.007</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>24</td>
<td>.31</td>
<td>.069</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>33</td>
<td>.22</td>
<td>.085</td>
</tr>
<tr>
<td>Winter 1988</td>
<td>18</td>
<td>.38</td>
<td>.058</td>
</tr>
</tbody>
</table>

Correlations between MTEL and TOEFL scores were somewhat higher than between the TWE and either of the two objective tests, but considerably lower than correlation figures claimed by the English Language Institute (Michigan Test Manual), as shown in Table IX.

**TABLE IX**

**MTEL/P/TOEFL CORRELATIONS**

<table>
<thead>
<tr>
<th>Term</th>
<th>Cases</th>
<th>Pearson r</th>
<th>Significance (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 1987</td>
<td>23</td>
<td>.23</td>
<td>.148</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>16</td>
<td>.52</td>
<td>.020</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>28</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>Winter 1988</td>
<td>15</td>
<td>.56</td>
<td>.015</td>
</tr>
</tbody>
</table>

The low correlations could be a result, in part, of the small sample size. In contrast to TWE/TOEFL and TOEFL intratest correlations, the greatest agreement between the MTEL and TOEFL was achieved in Fall, 1987. This is not the
only inconsistency noted in the analysis of MTELP scores. Although ELI maintains that scores for various forms of the test are equivalent, the disparity evidenced in first and second term test scores contradicts this claim. In three cases, first term MTELP scores were 30, 32, and 35 points higher than second term scores, a significant difference on a test with a ceiling score of 100.

**TWE/GPA Correlations**

The final question posed in this study is the predictive validity of the TWE. Can academic performance be predicted from English language proficiency test performance?

The score from the last TWE taken prior to exiting ESL classes was correlated with GPA in the first and second terms of academic coursework. The correlation between TWE scores and GPA was erratic, first-term GPA showed a very low correlation with TWE scores, whereas second-term GPA correlated highly with TWE scores, as revealed in Table X.
**TABLE X**

**TWE/GPA CORRELATIONS**

<table>
<thead>
<tr>
<th>Term</th>
<th>Cases</th>
<th>Pearson r</th>
<th>Significance (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TWE/FIRST TERM GPA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter 1987</td>
<td>11</td>
<td>.29</td>
<td>.196</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>10</td>
<td>.004</td>
<td>.495</td>
</tr>
<tr>
<td>Fall 1987</td>
<td>8</td>
<td>-.17</td>
<td>.340</td>
</tr>
<tr>
<td><strong>TWE/SECOND TERM GPA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter 1987</td>
<td>12</td>
<td>.75</td>
<td>.002</td>
</tr>
<tr>
<td>Spring 1987</td>
<td>10</td>
<td>.74</td>
<td>.007</td>
</tr>
</tbody>
</table>

The extremely small sample size is a result of students taking classes on a Pass/No Pass basis, which appears to be common during the first year of academic coursework. Added to this is the fact that some students did not continue at PSU, or did not enroll for a sufficient number of academic courses to be included in the analysis. All first-term correlations are below the significance level. The very strong correlations for the second term GPA are based on very limited samples, but are well above the significance level.

Correlations between the TOEFL and GPA, as illustrated in Table XI, proved to be as weak and erratic as those for the TWE, and above the significance level in all cases.
Correlations between the MTELP and GPA were as variable as those for the TWE and TOEFL. Correlations between Level 4 Michigan Test scores and first-term GPA yielded a negative correlation of .34, while second term GPA yielded a positive correlation of .42 (see Table XII).

TABLE XII
MTELP/GPA CORRELATIONS

FIRST TERM GPA

<table>
<thead>
<tr>
<th>Pearson r</th>
<th>Cases</th>
<th>Significance (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-.34</td>
<td>.168</td>
</tr>
</tbody>
</table>

SECOND TERM GPA

<table>
<thead>
<tr>
<th>Pearson r</th>
<th>Cases</th>
<th>Significance (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>.42</td>
<td>.148</td>
</tr>
</tbody>
</table>
All three tests showed a stronger correlation with second term GPA than with first term GPA, however, sample size was so small that all correlations were below the significance level. It was found that students often had lower grade point averages in the second term of post-ESL study (See Table XII), and this could be the reason for the stronger correlation.

Light, Xu, and Mossop (1987) found no significant correlation between TOEFL and graduate level GPA, but did find TOEFL to be a predictor of graduate credits earned. The correlation between TWE scores and credits earned was therefore investigated as a different measure of academic success (see Table XIII). Grade point averages were multiplied by the number of course credits completed to determine the number of credit points for each of the 21 students enrolled in graded coursework following ESL classes at PSU. Scores are ordered from most successful to least successful, based on both GPA and credits earned. All students carried a minimum of 8 credit hours.

Of the 12 students who completed 30.00 credit points per term (including the graduate student who achieved 29.97 points), 50% of them achieved a TWE score below 4.0, including one student who completed 20 credit hours in his first term of academic study. The graduate student who scored a 2 on the TWE successfully carried a full
# TABLE XIII

**TWE SCORES AND CREDIT POINTS EARNED**

<table>
<thead>
<tr>
<th>First Term Credit Points</th>
<th>Second Term Credit Points</th>
<th>TWE Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>56.00</td>
<td>32.00</td>
<td>3.5</td>
</tr>
<tr>
<td>52.00*</td>
<td>46.00</td>
<td>5.0</td>
</tr>
<tr>
<td>42.00</td>
<td>47.58</td>
<td>4.0</td>
</tr>
<tr>
<td>P(12)</td>
<td>36.00</td>
<td>4.5</td>
</tr>
<tr>
<td>31.50</td>
<td>29.25</td>
<td>5.0</td>
</tr>
<tr>
<td>39.96</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>39.27</td>
<td>30.00</td>
<td>5.0</td>
</tr>
<tr>
<td>39.90</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>33.00</td>
<td>29.97(G)**</td>
<td>2.0</td>
</tr>
<tr>
<td>31.92</td>
<td>16.00</td>
<td>3.0</td>
</tr>
<tr>
<td>30.00</td>
<td>15.00</td>
<td>3.0</td>
</tr>
<tr>
<td>29.97(G)**</td>
<td>8.72</td>
<td>2.0</td>
</tr>
<tr>
<td>28.92</td>
<td>26.10</td>
<td>3.5</td>
</tr>
<tr>
<td>27.00</td>
<td>P(3)</td>
<td>3.0</td>
</tr>
<tr>
<td>25.92</td>
<td>12.00</td>
<td>3.0</td>
</tr>
<tr>
<td>24.00</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>15.00</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>10.00</td>
<td>0.00</td>
<td>3.0</td>
</tr>
</tbody>
</table>

* Represents 20 credit hours with a GPA of 2.60.

** Represents 9 graduate credit hours with a GPA of 3.33.
courseload. Among the nine students who achieved less than 30.00 grade points, all scored 4.0 on the TWE.

It is difficult to generalize from such a small sample; however, it appears that among these students, those with TWE scores above 4.0 tend to perform better in academic coursework; however, a TWE score below 4 does not necessarily prevent a motivated student from successfully completing a full courseload.

SUMMARY

Interrater reliability for the TWE ratings at PSU appears to be high on average (.75), but below reliability figures claimed by ETS (.82 and .85). The lack of interrater controls may be one factor affecting scoring reliabilities.

TWE evaluations and instructors' rankings yielded moderate positive correlations (.39 to .50). The fact that correlations between the two were not stronger could be a result of the first-draft nature of the essay test, topic sensitivity, and rating inconsistencies. Instructor evaluations could also be responsible, in part, for the relative weakness of the correlations. Although instructors are exposed to a larger body of student work, they may not be free from bias in judging that work.

The low to moderate correlations between the TWE and TOEFL (.19, .36, .42, and .46) lend support to the view that
the two measures of English language proficiency are not evaluating the same skills. According to Oller (1978) and Carlson and Camp (1985), if correlations between integrative and discrete-point tests are strong, both are more likely to be valid measures of language proficiency. No figure has been given, however, for the degree of correlation to be expected between the two test types.

As claimed by ETS, intertest TOEFL correlations were high, ranging from .68 to .85 in this study. The low correlations exhibited between TOEFL Section 2, the structure and writing portion of the test, and the TWE (.06, .30, .34, .40), suggest that the two tests tap different skills.

Correlations between the TWE and the MTELP were similar to those achieved between the TWE and TOEFL (.22, .31, .38, and .48), but the different forms of the MTELP appear to yield disparate results, and MTELP and TOEFL correlations were far below reliabilities quoted by ELI (.80 to .94).

The number of students who continued at PSU, and whose graded coursework met the requirements of the study (an 8-credit minimum and restrictions on the kinds of courses to be included) was small, a fact which affected the significance level of the correlations between test scores and GPA. The only correlations above the significance level were those between the TWE and second term GPA. These
correlations were strong (.74 and .75), but based on such a small sample (10 and 12 students), and in such strong contrast to the other TWE/TOEFL and MTELP/GPA correlations, that they must be questioned.

It was found that neither GPA nor credits earned could be accurately predicted from TWE scores in a sample of this size (21 students).
I have reviewed some of the problems inherent in language proficiency testing, including restricting testing to language skills. When testing is integrative, i.e. requiring production of discourse, isolating language and cognitive skills is further complicated (Angelis 1977). However, if a proficiency test is to be used in the college or university admissions process, perhaps the isolation of language proficiency from intelligence is not crucial, since a student requires a certain degree of each to succeed in college.

Writing instruction today is process oriented. It is through writing that the writer learns what he has to say (Murray 1980; Taylor 1984). Often ideas do not arrive on time, in order, or neatly packaged (White, E. B., 1979; Zamel 1983; White 1986).

Essay tests favor those who extemporize well, and do not tell us who revises well. Another weakness of this kind of test is that it reflects only one mode of discourse, and each of us writes better on certain topics than others (Lloyd-Jones 1987). Objective tests, on the other hand, do
not inform us about students' organization skills (Mullen 1980; Odell 1981; Jacobs et al. 1981; Bamberg 1982; Benton and Kiewra 1986), and this information is particularly important because rhetorical strategies of non-native speakers often lag behind mechanical skills due to cultural differences in rhetorical patterns (Kaplan 1980; Carlson et al. 1985).

Essay tests of writing ability are complicated primarily by two major factors: the use of a one-session first-draft sample for evaluative purposes (Zamel 1983; White 1986; Carlson and Bridgeman 1986; Lloyd-Jones 1987) and the accurate analysis of that sample (Diederich, et al. 1966; Cooper 1977; Charney 1984; Homburg 1984; White 1986). Added to this are the problems of topic selection and topic equivalency (Jacobs, et al. 1981; Carlson and Bridgeman 1986; Lloyd-Jones 1987). These problems are particularly pertinent to this study, since topics used in the PSU version of the TWE test are still considered experimental. The moderate to low correlations between the TWE and evaluations of writing instructors are probably due in large part to these limitations.

The reliability of the holistic scoring of the TWE was reasonably high, .75; but scoring is one more factor to be added to an already variable writer's performance. An increase in scoring reliability might strengthen other
correlations. But even if raters are in agreement, this does not ensure that ratings are accurate.

Low to moderate correlations between the TWE and the total TOEFL score indicates that the TWE does to some extent reflect overall language proficiency; however, this correlation was not strong enough to confirm the validity of both tests. The conclusion could also be drawn that students who are better test-takers, regardless of skill, will outperform those who are poor test-takers on any test.

Michigan Test scores, while yielding correlations with the TWE similar to those between the TOEFL and TWE, showed evidence of inconsistencies among various forms of the test.

Studies investigating the relationship between TOEFL scores and GPA have yielded generally low correlations (Sharon 1971; Harvey 1979; Light, Xu, and Mossop 1987), and strong correlations between the TWE and GPA were not expected. The small sample size yielded correlations for all three tests which were not statistically significant for first-term GPA and questionable for second-term GPA. The analysis of credit points earned per term (derived by multiplying GPA by credits) revealed that students who achieved a 4.0 or above on the TWE were successful in completing academic coursework; whereas only 50% of the students with TWE scores below 4.0 achieved 30.00 credit points or above.
An issue to be considered in the analysis of GPA is course equivalency. Intuitively, one would assume that a mathematics course would make fewer demands on a student's English skills than an English literature or writing course. If the analysis of GPA is conducted in an attempt to demonstrate a relationship between English proficiency and academic performance, courses should be segregated by college or by department, but even this delineation would not guarantee course equivalency. Light, Xu, and Mossop (1987) found significantly higher TOEFL/GPA correlations for graduate students in humanities, fine arts and social sciences ($r = .24$) than for science, math, and business majors ($r = .04$); however, undergraduate coursework is much less specialized in nature than that of graduate students. Students must fulfill general university requirements, and often these courses take precedence over courses in the student's major during the first year or more of study. At Portland State, and probably at other colleges and universities, it is popular to take elective courses on a pass/no pass basis, and this makes the collection of GPA data difficult.

The relationship between language proficiency and academic success is a murky one, as noted earlier. Research into this area must take into account motivation, a very strong factor in academic performance and a difficult one to measure (Cortes 1967, Graham 1987). There does appear to be
a certain minimal language proficiency level, below which it is unlikely that a student will succeed. From the analysis of TWE scores and GPA in this sample it is difficult to determine what this level is. There are no students who performed poorly with a TWE score of 4 or higher; however, there were students who performed very well with a score as low as 2. Further evidence must be gathered to discover the frequency of success among students with low TWE scores.

CONCLUSION

This study investigated the validity and reliability of the Test of Written English, a recent addition to the TOEFL test. The Educational Testing Service considers the Test of Written English to be an experimental test and continues to monitor its use. Portland State University is assisting in the field-testing of TWE topics; as a result, topics used in PSU testing may show greater variation in student performance than those used in standard test administrations.

In addition to possible topic weakness, scoring reliabilities at PSU did not meet ETS levels. Under standard test conditions, TWE essays are scored at a centralized ETS location, rather than at individual institutions. Scoring accuracy at PSU may improve as raters gain more experience. ETS requires that raters meet together under the aegis of a Table Leader who periodically
checks ratings for uniformity. This kind of rating session is perhaps not feasible at PSU, but might enhance intrarater reliability.

Scoring categories for the TWE may provide another obstacle to interrater reliability. It is possible that the design of the rating scale overemphasizes grammatical correctness, or assumes that a certain level of mechanical and rhetorical competence will coincide, when, in fact, the disparity between the two aspects of writing ability may be great. Does the rating scale play fair with the more proficient writer who is willing to take syntactic risks in order to express him/herself? One experienced writing instructor commented that she found the TWE categories arbitrary, and found them difficult to work with.

Teacher evaluations of student competence did not demonstrate strong correlations with TWE scores. This lack of agreement could be attributable to several factors: (1) first-draft vs. writing with opportunity for revision; (2) test anxiety; (3) topic sensitivity, i.e. inability to write well on a specific topic; (4) topic bias (one chart/graph topic appears to favor geography students); (5) TWE rating guidelines; (6) instructor guidelines; (7) instructor bias; (8) test conditions; and (9) writer variation as a result of ill-health, etc.

Correlations between discrete-point and integrative tests should be strong if both are valid tests (Oller 1978;
Carlson and Camp 1985); however, correlations between the TWE and TOEFL/MTELP were low to moderate. Perhaps the validity of all three tests should be challenged, since the TOEFL and MTELP can only test language usage (formal rules) and not language use (communication) (Krashen 1982). Nonetheless, the results of this portion of the study confirm the hypothesis that the two types of tests measure different skills. Differences in scores on various forms of the MTELP indicated problems with intertest reliabilities.

GPA correlations for the TWE, TOEFL, and MTELP did not indicate predictive validity on the part of any of the three tests; however, sample size was insufficient for this part of the study, largely as a result of student attrition and the pass/no pass option. Three students achieved first and second-term GPAs above 3.00 with TWE scores below 4. No student with a TWE score of 4 or above achieved less than a 2.66 grade point average in the first-term of post-ESL study. Generalizations cannot be drawn from a sample of 21 students. If this were a representative sample, a required TWE score of 4 for undergraduates and 5 for graduate students would exclude both potentially unsuccessful students and potentially successful ones. In any case, it is extremely difficult to perform meaningful GPA calculations due to a popular Pass/No Pass option, differences in course load, and course and instructor
variability. A formula which takes these factors into account is necessary if GPA studies are to be definitive.

Grade point average is an important measure of academic success for those students who wish to enter graduate school; for others, completing college coursework in a timely fashion is of greater importance. Studies examining GPA should also take into account credits earned. The results of the grade point analysis which combined GPA and credits completed indicated that students who achieved a TWE score of 4.0 or above were successful in their academic coursework; whereas the remainder were as likely to fail as to succeed.

If TWE scores are adopted by PSU as indicators of academic readiness, a cut-off score of 4.0 would probably eliminate many potentially unsuccessful students. Perhaps for students who achieve scores below 4.0, other factors should be weighed against TWE scores, factors such as instructor recommendations, objective test scores, and previous academic records.

In conclusion, any assessment of language proficiency is affected by confounding factors which distort results. No one instrument should be given too much weight in the admission process. It should be remembered that academic success is not determined solely by language proficiency, but is heavily influenced by other factors, particularly academic aptitude and motivation.
IMPLICATIONS

The results of this study and the questions raised by the analysis suggest several implications for further research.

One possible area of investigation is the question of first-draft writing samples. A comparison of first- and second-draft samples would provide a clearer picture of writing competence.

Research in holistic rating could focus on a comparison of holistically-scored essays and essays given separate scores for organization, syntax, appropriate use of vocabulary, and other features. Non-holistic systems are not necessarily complex; Jacobs, et al. (1981) have designed a profile system for essay scoring which gives separate scores for five categories, but which can be used efficiently. A study involving both scoring methods would provide additional information concerning the reliability of the holistic scoring.

Holistically scored essays could also be correlated with scores achieved on sentence-combining and sentence-and paragraph ordering tests as described by Gaies (1976).

PSU has relied heavily on compare/contrast topics for the essay test, although many international students pursue studies in the sciences where interpreting a chart or graph may be of equal or greater importance. Studies comparing performance on both topic types would yield additional
information on topic influence in test performance. Although the Educational Testing Service has conducted such studies (Carlson, et al. 1985), results at PSU would not necessarily mirror those attained by ETS.

There appear to be significant variations in test performance on various forms of the Michigan test. Since this test is in active use at PSU and plays a role in admissions decisions, research should be undertaken in this area.

Although GPA studies are complicated by several factors, institutions of higher education need to make decisions regarding international student readiness for academic coursework, and GPA is one accessible measure of academic performance. Studies involving GPA should be longitudinal, since sample size is quickly narrowed if restrictions are put on the number or types of courses to be included in the analysis. The number of credits completed per term represents another aspect of student achievement, and should be combined with grade point averages.

Although the ability to write clearly in English is an important factor in an international student's future success, neither the measurement of writing competence nor that of overall language proficiency is exact. Collecting as much data on language proficiency as possible will increase the chances of making accurate admissions decisions. English proficiency may not necessarily be the
deciding factor in a student's academic performance; however, required levels must be set high enough to minimize frustration and failure on the part of both students and faculty.
REFERENCES


APPENDIX A

THE TOEFL TEST OF WRITTEN ENGLISH (TWE)

Examinees are told that they will have 30 minutes to plan, write, and correct an essay on the assigned topic and that they should write clearly, use examples to support their ideas, and check their work in the last few minutes.

TWE readers assign scores based on the following scoring guide. Though examinees are directed to write on the assigned topic, parts of the topic may be treated by implication.

TWE SCORING GUIDE

SCORE

6  Clearly demonstrates competence in writing on both the rhetorical and syntactic levels, though the essay may have occasional errors.

A paper in this category:
- is well organized and well developed
- effectively addresses the writing task
- uses appropriate details to support a thesis or illustrate ideas
- shows unity, coherence, and progression
- displays consistent facility in the use of language
- demonstrates syntactic variety and appropriate word choice

5  Demonstrates competence on both the rhetorical and syntactic levels, though it will have occasional errors.

A paper in this category:
- is generally well organized and well developed, though it may have fewer details than does a 6 paper
- may address some parts of the task more effectively than others
- shows unity, coherence and, progression
demonstrates some syntactic variety and range of vocabulary
displays facility in language, though it may have more errors than does a 6 paper

4 Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.

A paper in this category:
- is adequately organized
- addresses the writing topic adequately but may slight parts of the task
- uses some details to support a thesis or illustrate ideas
- demonstrates adequate but undistinguished or inconsistent facility with syntax and usage
- may contain some serious errors that occasionally obscure meaning

3 Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level, or both.

A paper in this category may reveal one or more of the following weaknesses:

- inadequate organization or development
- failure to support or illustrate generalizations with appropriate or sufficient detail
- an accumulation of errors in sentence structure and/or usage
- a noticeably inappropriate choice of words or word forms

2 Suggests incompetence in writing.

A paper in this category is seriously flawed by one or more of the following weaknesses:
- failure to organize or develop
- little or no detail, or irrelevant specifics
- serious and frequent errors in usage or sentence structure
- serious problem with focus
Demonstrates incompetence in writing.

A paper in this category will contain serious and persistent writing errors, may be illogical or incoherent, or may reveal the writer's inability to comprehend the question. A paper that is severely underdeveloped, or one that exhibits no response at all, also falls into this category.

Papers that reject the assignment, fail to address the topic in any way, or exhibit absolutely no response at all must be given to the Table Leader.

Copyright ETS 1987
APPENDIX B

Level 3 and Level 4 Writing Instructors:

Please assist me in my research by rating your students' writing ability according to the following guidelines. Thank you for your help.

1. A rating of 1 indicates that the student is incompetent in writing. This writer fails to use a topic sentence, doesn't provide support or details where necessary. Overall organization is poor or non-existent. Grammatical and lexical errors are frequent and often obscure meaning.

2. A rating of 2 indicates that the student fails to organize essays. There is a general lack of supporting details, while extraneous ideas and statements are included. Mechanical errors are frequent and make interpretation difficult.

3. A student with a rating of 3 is beginning to develop some writing competence. Organization and/or development are inadequate. Details and supporting evidence are given, but incomplete. Errors in sentence structure and/or inappropriate choice of words are common.

4. Writing which merits a rating of 4 is nearly competent. Papers are adequately organized and details are used to support the thesis. There is some use of more complex structures. The student does not yet demonstrate a broad range of vocabulary and syntax. Writing seems cautious; the student is avoiding useful structures in order to make fewer grammatical errors.

5. A rating of 5 represents writing competence, both syntactic and rhetorical. Essays are well organized and developed. Vocabulary usage and sentence structure are varied and appropriate. There are few, if any, inconsistencies and omissions in supporting details. There are few errors and these generally do not obscure meaning.
Writing at the 6 level displays consistent competence in English. A student at this level would have no difficulty in producing a well organized, well developed essay which effectively addresses the writing task. Syntax and vocabulary errors are minimal.
APPENDIX C

TEST OF WRITTEN ENGLISH

ESSAY TOPICS

Applicants at seven U.S. colleges and universities wrote on the following eight topics during TWE pilot-testing in 1983. (Carlson, et al. 1985). Sample essays written on two of the topics are given in Appendix C.

Chart/Graph

1. Individual consumption of major foods in the U.S. (line graph).

2. Changes in automobile part production by three companies (bar graph).


4. Area and population of continents (two pie charts).

Compare/Contrast

1. Travel and reading are two ways of learning about people and the world.

2. Advantages and disadvantages of exploration of outer space.

3. Advantages and disadvantages of using chemicals to control insects.

4. Preference for spending free time in active, physical recreation vs. participation in intellectual activities.
APPENDIX D

TWE SAMPLE ESSAYS

The following sample essays, representing each of the scoring categories for both topic types, were provided by Educational Testing Service.
Some people say that exploration of outer space has many advantages; other people feel that it is a waste of money and other resources. Write a brief essay in which you discuss each of these positions. Give one or two advantages and disadvantages of space exploration, and explain which position you support.
Scientists make great effort in space searching. It is a great thing that human being can send the boat into space and advance search is continued. To conquer the space in the most interest thing that humans want to prove.

Both great nations U.S.A and U.S.S.R. have a competition in space exploration for many years but what have they really got in search is quite doubtful. All we people have seen nothing but space man walking on outdoor spaceboat and the pictures with Mars and Moon. Have they ever found an E.T.? Spent a lot of money in exploration outer space but got few that seems very stupid action that we have done. How many years they can have a result in our own eye with extrasppe travel? The scientist is not sure for these question. It has argument in Congress if we should spent some money in exploration. Eye on the advantage of space exploration. We can make it possible — live in the outer space and travel in the space, even make friends with outer space people. So why not have patient on searching space?
An interesting idea to talk about is the topic of outer space especially the exploration of outer space. However, there are not enough information in my study regime. I could express my thought with my text.

The common sense is widely known that the solar population in the small earth became and the larger people find the higher technology developed to the resources and situation now cannot satisfy.

Meanwhile, if the people on the earth would like to devote themselves to explore the outer space, they would not pay much attention to manifest the weapons to destroy the human kind.

Though the exploration of outer space has many disadvantages such as money and time, we should show that whatever you do, you need time and money. And why shouldn't we spend our capacity to explore the something new for the long life of human kind. There are many unexplored situations like resources, radiation, nature lost, dirty water, air pollution and garbage like mountains. That means something we cannot neglect. Why don't we find something new and available for the human kind? (continued)
I would like to express my point of view that I prefer exploration of outer space to not exploration of outer space. It is because the population grows at a faster rate than every living thing on the earth. Therefore, it may be true that only after several years, there will be famine, lack of living space, and so on. The problem of pollution inevitably arises from the proliferation of the industrial side. To breathe a piece of fresh air is no longer an easy task. Moreover, we should be noted that although the improvement in technology is very prominent, it cannot really "create" some natural resources of food, talking at a greater extent. The land can be increased by means of reclamation; however, it is very very costly and it is not easy to do so. As a consequence, exploration of outer space we can be able to discover more areas for development and at the same time to solve the problem of the exhaustion of natural resources.

However, nothing will be perfect.
Despite the fact that it can favour new development into the outer space, it would take a tremendous amount of money to install all the facilities and also to employ the top-grade scientists. Some poor blockade may even result at the environment and out of them why not to spend more money into the activity in the form of welfare payments to reserve their living standard or some other business men may wonder why the government spend so much money on the "living" activity. Therefore, we can see that if there is no link at all in the exploration of outer space, everybody will clap their hands because this will enhance the living standard very much, however, in reality, nothing will be certain and the different opinions arise from different walks of life.

Finally, I would like to stress that although I support the exploration of outer space, I don't mean that it is perfectly good for all.
Outer space has always been a mystery for curious human beings. With the invention of the microscope and its development, we have found out many things about the universe that surrounds us, and we still have a long way to go. I think that exploration of outer space is fascinating; there are still so many things to discover that we can't even imagine! I belong to the group of people who think that it is impossible that in such a big universe we would be alone. I just can not believe such a statement. My opinion is that we should be as open-minded as we could because there are still so many things we don't know, but day by day we learn a little bit more.

Another advantage of exploring outer space is that if there would be another war that would destroy the world, or if we were too many and one of us could try to take off and begin a new life in another or maybe even in a distant planet. I think that would be difficult but very exciting!

A disadvantage of exploring in outer space is that a lot of money has to be invested in the construction of rockets, satellites, and in the preparation of the staff that works in such a complex. And I see this as a disadvantage because with that money many starving people could be fed.

I support the exploration of outer space because I think it to be very important for the future when this world won't have a thing to offer to its inhabitants.
Some people say that exploration of outer space has many advantages. Firstly, it leads to the rapid development of advanced scientific technology. From the first landing on the moon in 1969, the development of space shuttle, it only costs the scientists several years. It is really a great accomplishment with such advanced technology. Many living standards can purely be raised. Harvard Exploration of outer space may also help to solve the population problem on the earth. With the success of space shuttle migration to other planets in the future is not impossible.

However, there are people criticizing the exploration of outer space. They feel that it is a waste of money and other resources. Every year, billions of dollars are spent in developing space rockets, shuttles and other technologies. If this amount of money is used in social welfare or is used to help those who are hungry in the poor countries, many people's lives can be much easier. Besides, the exploration of outer space also leads to the development of space weapons. These weapons, when in case of war may cause a complete destruction of the earth. Therefore, exploration of outer space is disastrous to human beings.
In my opinion, although there are many disadvantages of exploration of outer space, I personally feel that it is worth doing. "Curiosity is the mother of invention." People are curious of the earth, and therefore, learn more about it. However, if they knew that the earth is round and is one of the numerous planets in space, at present, we do not really understand outer space. Therefore, we have to learn more about it. Otherwise, there will be no improvements in our society. After many years of exploration, we may contact another civilization. This is very challenging. Therefore, exploration of outer space, though it is quite risky, will definitely benefit us because we will learn more and more from the dark space.
Outer space has always been a mystery to us, and
it has been the heated topic for discussion among not
only scientists but many other people of different walks.
To know more about outer space will mean more
exploration of outer space, which may require a lot of
money and resources. But is outer space really worth
exploring?

One of the aims to explore outer space is
to find out whether other forms of life exist in places other than
the earth. People have always been curious about the
question that extra-terrestrial exist; to explore the outer
space may possibly enable to find out the answer. Should we
really find some other kinds of organisms, we may benefit
from them because our level of knowledge may be
raised as a result of communication with them.

Secondly, exploration of outer space may enable
us to discover some new elements or other scientific laws.
This is quite true as the chemical element helium was first
found when the moon was observed by a scientist. Thus,
if we continue to explore outer space, it may be
likely that we can improve our scientific techniques
and acquire more knowledge about the universe.

However, although exploration of outer space has
several advantages, it costs us a lot of money and
resources. Millions of dollars have been used for building
a single rocket, and billions of liters of liquidized oxygen
and nitrogan have been used for fuels. How much
money and fuel can we save if the resources spent
on exploration are used for people of different countries!
We are now, moreover, facing an energy shortage. Should we continue to waste that incredible amount of energy on exploration of outer space?

Furthermore, exploration of outer space has diverted a lot of scientists and labor from other departments. The exploration seems to help only those super-countries, while the underdeveloped countries will suffer an aggravated lack in human resources. As a result, the rich countries which have the ability to explore outer space will become more powerful, while those poor countries will remain in their conditions or even suffer further, because they cannot receive help from others who concentrate on exploration.

Thus, exploration of outer space has both advantages and disadvantages. Then which should we choose, to continue or stop at once the exploration?

In my opinion, there is the need to explore the outer space. Our life may be improved as a result of exploration, not only in the aspect of science. Perhaps we may discover an element or a technique to cure the illness of the people who live in the dirty conditions of the underdeveloped countries. Instead of spending our energy and resources in exploring the outer space.
Suppose you are to write a report in which you interpret these charts. Discuss how the information in the Area chart is related to the information in the Population chart. Explain the conclusions you have reached from the information in the two charts. Be sure the charts support your conclusions.

YOU MAY USE THIS SPACE FOR NOTES.
Population problems are getting important in today. Because population problems are released easy health. But air can cause your breath detect. So population problems are necessary to think. From the chart, we talk about me fellows.

You know, the Asian area is 30%, and its population is 59%. In other side, the Africa area is 20%, population is 11%. Only they have different situation, why? Maybe it indicated develop problem.

In other side, the Asia area is 20%, its population is 59%, and the North America area is 16%, its population is 97% only. They have different situation, why? If we use "develop" problem, its youth. Above 2 kinds of states, we know develop problem is not to cause population problem. North America have slightly develop, but its population is less than another.

So, we know, the important is quality. If you know "Do your best", population problem will reduce more. We will have a good and health world.
For more a report, we must analyse both charts continent by continent. As we can see, the population in Asia is more than a half of the world population, and this territory has the biggest rate, but it's only 30% of total area, that means there is a lot of people in one determinated area. In other words, the people is not homogeneously distributed. In the other hand we look for example Oceania which has 6% of total area and only 1% of world population, that let us to suppose that there is more space for few people, and also means heterogeneous distribution of population.

We can see also there is many territory in Africa for relatively few people and in Europe there is a lot of people for this little continent.

The conclusion about this data is the heterogeneous distribution of population all around the world which could cause a kind of desequilibrium. Another conclusion could be that in certains continents there is a high demographic growth which causes this bad distribution like as we can see in Asia and in the other side Oceania, Antarctica could not.
The land occupies three-fourths of the total area of the earth. In this small piece of land, there are over 200 countries and the five continents we list by Latin name. There are over 5 billion people living in the world.

Asia is the largest continent among the five, and it has the largest population. It occupies 30% of the total, whereas its population is 60% of the total. An outstanding population, it can be seen that Asia is densely populated. North America is the next largest after Asia because it has only one-third of the total population. Because of small population, the living standards of the United States and Canada are more advanced than other countries. Europe is a small continent with a high population density. It includes several small island countries in recent years. The world is in favor of countries like Japan. There are many countries in Europe which the world as Europe.

Africa is a very small continent. It has only 1/8 of the total population although Africa is a large continent. The population increased because of the increase of the activities of the forces of destruction. Many poor Africans were forced to behave like the blacks. South Africa is a country with a very small population that Europe although it has a large land. The population is so sparse that many of them are afraid to live there. It affects the growth of the South Africa.

From the above, we can see that North America is the most important among the five. Its technology, living standard, and resources are in the lead in the world. The most densely populated areas are the home of the living standard and the wealth of the nation.
Studying the information given, the area chart is related to the population chart in many ways.

First, the most obvious conclusion is that Asia has the largest area and the biggest population, and, even though this is true, we know that most countries in Asia, like China, have a very big problem of population explosion, therefore, the big area they have is not enough.

Europe has a small area, 7%, and it for the population of 16%, that means they double of the area, but still they don't suffer from what Asia suffers from.

North America has a large area, compared to the population, and they don't have a problem of population, as most people know to use methods of birth control.

As for South America, they have no problems as they have a small population and a large area, which most of it is still unexplored jungles, the same conclusions go for Africa.

Antarctica has a fairly good area 9%, but it has no population 0%. Most of Antarctica is icy, it has very poor weather conditions and the population that exists there doesn't add up to 0% percent. There are no people.
people who live there, but sometimes some research groups go there.

Oceania consists of islands all over the world, it also has a good area, but the population is one of the smallest, as there is a lot of islands spread out in the oceans, which are not inhabited.
The population around the world is not evenly distributed. The area with the highest density and the highest concentration of people is usually highly industrialized. This is because of the need for labor and resources. In the areas where the population is sparse, the quality of life is much lower. The quality of life in these areas is affected by the lack of resources and the absence of proper infrastructure.

Africa has one-fifth of the total area and 11% of the total population. We have to be careful with these figures, because some of Africa's surface is a desert, making the climate and resources for people sparse. But the lack of economic and natural resources have caused that there is little change in the area's habitat in more than 400 years.

Asia is the largest continent on earth, 30% of the total surface, but has almost 60% of the world's inhabitants. This striking fact has to be explained by the way people live in this continent. In the countries of Pakistan, India, Pakistan, Malaysia, etc., there are no ways to control the rate of birth, and no social services to help the people with their children. Of course, there are two of the largest countries in the world on this continent, China and the Soviet Union; the latter has to be taken in a separate way because it uses a more developed country, but China, besides of the same problem of the rest of these countries...
BEGIN WRITING HERE

Although they comprise only 37% of the world’s total land area, Europe and Asia make up 74% of the population. This implies a much greater population density on the wide, for countries on these two continents than for others. Antarctica, for example, contributes 0.1% to the total land area, but 0.9% to world population. Similarly, the countries of North & South America, Africa, and the regions labelled Ocean, each support much lower populations than might be expected if land area is the sole factor controlling distribution. Obviously, there must be other criteria which influence human distribution.

Not all of the land’s surface is amenable to development and settlement by mankind. The world’s mountain ranges, for example, are too hostile to allow major population centers to develop. While it is not impossible to live there, they can only support a limited number of people. Another inhospitable type of surface would be desert. Once again, due to adverse conditions, it is hard for people to live in large numbers in the desert.

Climate is a factor which should not be neglected in the analysis of this problem. It was alluded to in deserts, as they are in part, climate-induced. However, there are large areas which, even if they did not count as desert (or arid), would be
uninhabitable. I refer, of course, to Greenland and Antarctica. Both of these are in the
southern sense, deserts, neither receive such precipitation in the course of a year; however,
the stronger influence in these areas is temperature. They are each too cold to
support the complex array of plants & animals needed in order to flourish.

Food supply, and some raw materials,
are also linked both to climate and to human
population density. Where conditions favor the
development of abrided nutritional sources
and raw materials such as wood, man is likely
to be found and in great numbers. In areas
where these commodities are scarce, man will not
be so evident.

While these are not the only parameters
which control the distribution of human populations,
accessibility, climate, and availability of the basic
necessities of life are important considerations.
On continents where these coincide favorably man
has settled in great numbers. Where there are
not as well-developed mankind tends to be
more thinly distributed. Europe and Asia seem
to have the closest to optimum combination
while Antarctica appears to be uninhabitable.