CANADIAN THESES

THÈSES CANADIENNES

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30.

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ MICROFILMÉE TELLE QUE NOUS L'AVONS REÇUE

An Analysis of Testing Measures Used to Screen Applicants for Admission to University

John Russell Scollan

A Thesis

in

The Centre for Teaching English

as a Second Language

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montréal, Québec, Canada

December 1986

ISBN 0-315-37112-9

ABSTRACT

An Analysis of Testing Measures Used to Screen Applicants for Admission to University

John Russell Scollan

Holistic scores given to compositions were examined to determine the amount of error due to intrareader reliability, interreader reliability, and to determine if the scoring of compositions is affected by whether or not the writing of the students is difficult to read. The scores of the compositions were also compared to scores on a multiple-choice test written by the same students to see to what extent the reliable variance in one test can be predicted by the reliable variance in the other. It was found that intrareader reliability and handwriting legibility did not significantly affect scores, but it was found that who the reader is significantly affected score magnitude. It was also found that approximately half the variance in composition scores could be predicted by the scores on a multiple-choice test.

Table of Contents

FIGURES

TABLES

Chapter One

Introduction

At Concordia University, Montreal, the English language ability of students who speak English as a second language is evaluated by a battery of two tests: the Michigan Test of English Language Proficiency (MTELP), a multiple-choice test of general English proficiency, and a Pragmatic Task (PT). To complete the PT, an applicant must read a short argument and counter-argument on a given topic. Then the applicant must take a stand and write a composition using his own ideas and incorporating those just read.

The Rationale for Having a Pragmatic Task

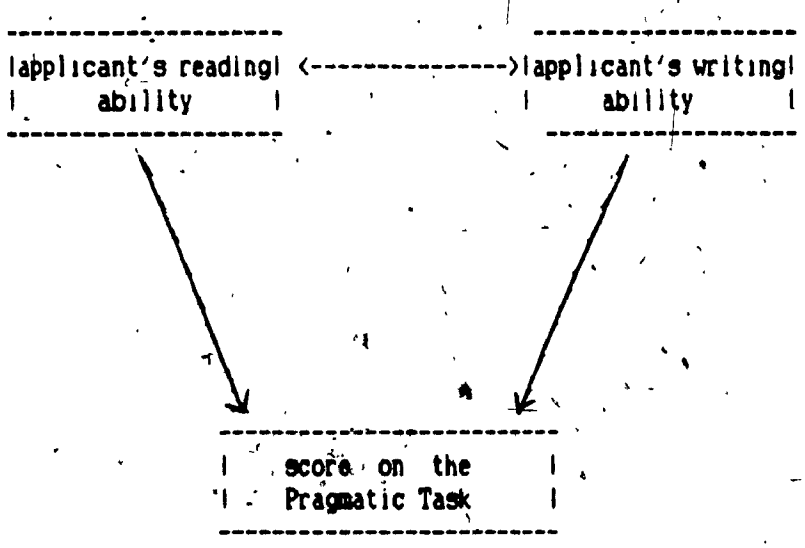
The PT requires the applicant to perform a task which is, to some extent, representative of the actual type of work he will have to perform as a university student, that is, reading something in English and writing something which addresses itself to the material which he has just read. The MTELP is an index of the level of an applicant's knowledge of the English language. The PT is used to determine how well he can apply this knowledge. That is, in the PT the applicant can read a short text and he can write a

composition in which the ideas just read are incorporated with his own to make an argument. It is therefore intended that the PT give us information concerning an applicant's English language abilities that is not available in the information that is obtained from the MTELP.

Ideally, an applicant's score on the PT is an accurate reflection of his reading and writing ability in English. This relationship is shown in the causal relation in Figure 1.

Figure 1

Ideal Relation between PT Scores and the Language Abilities Being Tested.



Purpose of the Study

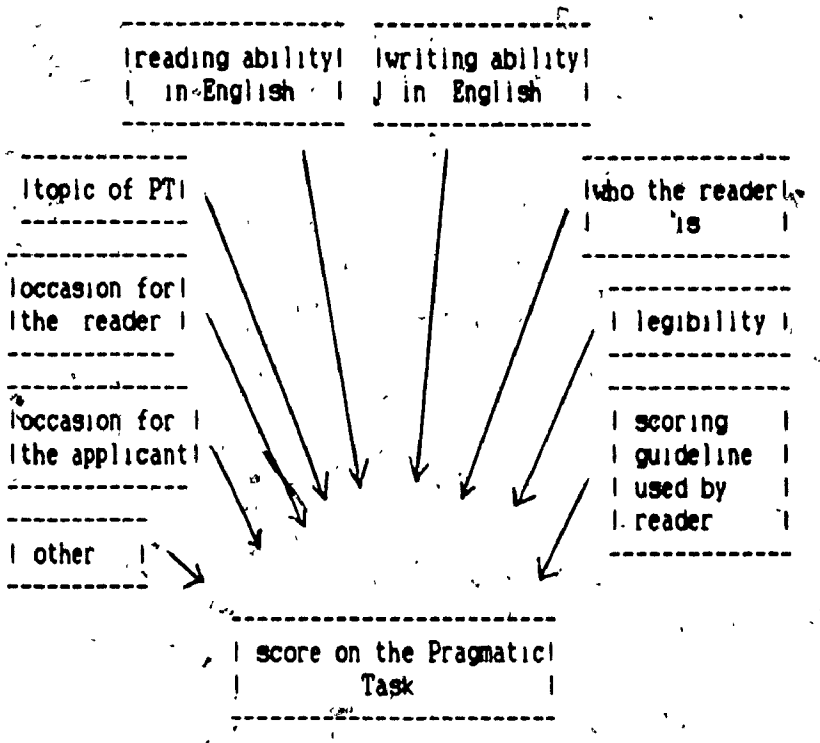
The study is designed to investigate the extent to which PT scores are independent of MTELP scores and the extent to which this independence is attributable to factors other than writing ability in English.

Factors other than reading and writing ability which might affect an applicant's score and which are examined here are (a) occasion for the reader, (b) who the reader is, and, (c) the legibility of the applicant's handwriting. Other factors which can affect PT scores such as, (a) PT topic, (b) occasion for the applicant, and (c) scoring scheme are not examined in this study.

In psychological and educational testing it is never possible to have pure tests of traits or abilities. The PT is a testing instrument whose purpose is to give us an accurate index of an applicant's ability to use the English language in terms of reading and writing (Figure 1). In Figure 2 this relationship is conceptualized in more realistic terms.

Figure 2

Factors Affecting PT Scores (Interactions are not Indicated).



The factors shown in Figure 2 can be considered conceptually independent from one another. We can, for example, imagine the factor topic as being distinct from that of who the reader is. Furthermore, these and other factors can be defined operationally where the effect of a given factor on PT scores can be described. In fact, however, we can expect interaction to occur among the factors shown in Figure 2.

Factors Affecting Composition Scores

In this section the factors shown in Figure 2 (p. 4) are discussed in general terms. These factors are: (a) topic, (b) occasion for the reader, (c) who the reader is, (d) legibility, (e) occasion for the applicant, and (f) scoring scheme.

1. Topic (Effect of Type of Discourse)

Discussion on this factor can become confused due to various understandings on its meaning. From the literature it seems that some authors understand topic to mean different types of discourse. With this understanding, different topics could mean an argumentative composition versus a descriptive composition. Other authors simply mean different 'titles', where the type of discourse is the same. For example, with this approach one might understand the following to be two different topics: 'How I Spent My Summer Holidays' and 'How I Spent My Christmas.' For both meanings the factor topic can have an effect. Even where the type of writing required is the same, the title might have an effect on a reader, applicant, or both, due to interest or dislike. A Moslem applicant might happily describe his holiday but be

distracted by having to deal with the idea of Christmas which is foreign to his religion.

Therefore, the topic for a composition can affect an applicant's performance positively or negatively due to such factors as interest and knowledge. Given a different topic the applicant might perform better and receive a higher score. Similarly, topic can also affect the readers, because of their interests, knowledge and attitudes.

Beyond subject matter, different topics may require different English language skills. Some topics call for descriptive writing while others demand argumentation. An applicant might be better prepared to handle the skills required for one topic than those required for another.

The language skills required to deal with a given topic can also influence the reader. Some readers might focus primarily on errors and ignore the fact that a subject is attempting a wider range of structures, including some which are quite difficult. Another reader might be favourably impressed by a piece of writing which shows a wider range of grammatical structures despite the errors in grammar.

2. Occasion for the Reader

The occasion on which a reader scores a composition might affect the score given to the composition. If the score given a PT by a reader were a correct reflection of the reading and writing ability of the applicant, then even when the reader scored the same PT more than once, the score would always be the same. But if a reader gave different scores to the same PT at different times, then to some degree the score would be a reflection of the reader's inconsistency and not of the skill of the applicant in writing the PT. One question posed in this study is, 'To what extent would a reader's score change if he were to score the same compositions on two different occasions?' Such score differences could be examined in terms of score magnitude (means), dispersion (standard deviation) and ranking (correlation).

3. Who the Reader Is

The question here is whether or not different readers assign the same scores to a set of compositions. If the readers score the compositions similarly then who the reader is is not an important factor affecting composition scores. However, if it is

a factor, it might be having an effect in one or more of several possible ways.

To what extent and in what ways might two readers agree or disagree on how they score the same compositions? Do they give high and low scores in similar fashion? Do they cluster or spread their scores differently? The ranking of the scores they give a set of compositions might be identical with their respective means and standard deviations being notably different.

Two readers might consistently agree on which compositions are relatively good and poor (rank order), with one reader consistently giving higher or lower scores (mean). Furthermore, one reader might be more consistent in how he scores the better compositions as compared to how he scores the poorer ones. Some readers might distinguish well between good and excellent compositions while being unable to differentiate between weak and hopeless ones. Other readers might avoid high and/or low scores and so cluster them towards the middle.

Despite the use of a guideline in scoring, different aspects of English might be more important or influential with some readers than with others. Some might be more influenced by the overall look of a composition in terms of its interest, logic and

cohesiveness. Others might be more influenced by the mechanics such as grammar, vocabulary and orthography.

In other cases some readers might be influenced by the same factors but to different extents or in different ways. Different kinds of grammatical errors, for example, might be viewed as more or less important by different readers.

The upper limit of consistency that can be attained between two readers is limited by the consistency in scoring readers have with themselves (occasion for the reader).

If who the reader is has an important effect, then it represents an undesirable factor affecting essay scores.

4. Legibility

The legibility of an applicant's handwriting is a factor which may influence the score a reader gives on a composition. An otherwise excellent composition might receive a low score due to difficult-to-read handwriting.

Some readers might be more skilfull at deciphering some difficult-to-read handwriting and so be less influenced by this factor than other readers. In such a case legibility would be a factor affecting the variance in composition scores.

Other readers might tend to give an applicant the benefit of the doubt when unable to read the handwriting thus possibily giving a higher score than warranted. Some readers might be adversely affected by such writing and the visual struggle it confronts them with and so be prone to suspecting errors where legibility is the only real problem. In such cases some otherwise excellent compositions might be given low scores.

It is also possible that legibility plays a more complex role. Different types of handwriting might be more or less problematic with different readers. Some readers might gloss over possible errors to a certain point and then quickly become negative while others might change more gradually in how they are affected by legibility problems.

5. Occasion for the Applicant

The occasion on which an applicant writes a composition might have an effect on his performance and hence his score. If an applicant is rated on the basis of an averge number of compositions over a period of time, then occasion for the applicant is no longer a factor, but if he is being scored on the basis of his work on one occasion only, then it can be important. Furthermore some applicants might be more emotional

with periods of high and low productivity while other applicants might show less fluctuation in their day to day work. In the former case, a larger sample of an applicant's work would be necessary to get a clearer picture of his overall ability in expressing himself in English.

6. Scoring Scheme

Scoring schemes can differ greatly and have an effect on how a piece of writing is scored. Methods for scoring can vary from counting words and words per sentence to general overall impression. What is being considered here refers to holistic schemes only. Such schemes, which give different instructions to the readers, might have three, five, ten or maybe twenty point scales which act as scoring guidelines, but in the final analysis, the subjective impression of the reader determines the score. The question then is, do these scoring schemes, which are basically the same but nevertheless different, affect how essays are scored?

Conclusion

In this section the factors affecting composition scores and shown in Figure 2 have been discussed in general terms. Of these factors (a) legibility, (b)

who the reader is, and (c) occasion for the reader are examined empirically. Due to the limitations of this study the factors (a) topic, (b) occasion for the applicant, and (c) scoring scheme, which have been discussed in general terms, are not examined. However, the role these factors have been seen to play in other research is presented in the review of the literature.

The effect or variance in composition scores accounted for by the factors shown in Figure 2 may be systematic but still unintended and therefore undesirable, because their effect on PT scores is not determined by the English language skills of the applicants. The extent of this variance places a limit on the amount of variance that can possibly be due to reading and writing skills in English as shown in Figure 1. Once the amount of error variance is determined, the next question is how much of the remaining variance is independent of that accounted for by an objective multiple-choice test, which in this study is the MTELP.

In the following section the relationship between composition scores and objective multiple-choice tests is discussed in general terms.

The Independence of Composition Scores and Multiple-Choice Test Scores

In this section the relationship between a composition type exam such as the PT and an objective multiple-choice test is examined. The basic question of this study is, 'How much unique information concerning the English language abilities of the applicants does the PT give us?' To some extent PT scores are not a function of the language skills of the applicants. Instead, they reflect other factors such as topic and legibility which subtract from the information value of the PT scores obtained. Hence, a certain percentage or proportion of a score would be a result of error.

At Concordia University, applicants who write a composition type exam also write an objective multiple-choice test. The question then is, what proportion or percentage of the PT score variance not attributable to error is independent (not predicted by) the scores the same applicants obtain on the objective multiple-choice test. The proportion of PT score variance not predicted by an objective multiple-choice test and not attributable to error would represent the maximum amount of unique information such a test would be giving us. Of course, it is possible that other

factors causing error have an effect even though they are not examined nor discussed here and in the literature.

The purpose of this study then is to determine what percentage of PT score variance is not due to error and is not predicted by the MTELP scores obtained by the same applicants.

Summary of the Chapter

In this chapter the purpose of the study has been discussed in general terms. First, the ideal relationship between an applicant's English reading and writing skills and the score he obtains on a writing task (Figure 1) was examined. Next, sources of error in composition scores were shown in Figure 2 and discussed each in turn. Then the question of independence between composition type scores and objective multiple-choice test scores was discussed.

Finally, the purpose of this study was defined as determining the percentage of PT scores which is independent of MTELP scores and which is not due to error. These questions are important to both university applicants and the university itself. The purpose of these tests is to inform the university concerning the English language abilities of the applicants and so it is important to both applicants

and administrators that the tools being used to determine the English language abilities of the applicants be as error-free and informative as possible.

The following chapter is a review of the literature and looks at other studies which have examined questions similar to those which are dealt with in this study. Some of these studies have examined the same questions which are examined in this study and allow for direct comparisons. Others have examined aspects of the question which the design of this study does not permit us to look at. In these cases these other studies fill in some of the gaps left by this study.

Glossary

Applicant: A person wishing to be admitted to the university who writes the CELDT.

CELDT: Concordia English Language Diagnostic Test.

Format: In this study this term is used to contrast between PTs scored by readers in the handwriting of the applicants and those which were scored by the readers after they had been typed.

Holistic Scoring: K. Perkins (1983) writes, "Holistic scoring evaluates a whole text rather than simply parts of a text." Further he writes, "In scoring holistically the grader reads the composition, forms a general impression, and assigns a mark to that composition based on some standard" (p. 652, 653).

Reader: A person who reads a composition, essay or Pragmatic Task with the purpose of giving it a score.

Pragmatic Task: The Pragmatic Task (PT) requires a person to read something in English, and to write something which addresses itself to the material read. The ideas read must in some way be incorporated with the writer's own ideas to present an argument.

Score: The mark a reader gives an essay based on his or her perception of its merit. The word score was used instead of grade because it was felt that a grade might imply corrections with the view of helping the writer improve future essays.

Chapter Two

Review of the Literature

In this chapter the results of other studies are examined. The chapter is divided into 5 parts. Each part deals with a different aspect of the problems relevant to making tests of writing ability. They are: (a) topic, (b) occasion for the reader, (c) who the reader is, (d) scoring scheme and (e) objective and composition measures. The first four parts are concerned with possible sources of scoring errors and the last part deals with the question of the independence of essay-type scores and objective multiple-choice tests.

Legibility is not discussed in this chapter because, although it is part of the study, no other studies examining this question were found. The factor, 'Occasion for the Applicant,' although discussed in general terms in the previous chapter, was not looked for in the review of the literature nor was it examined in this study.

A study which deals with more than one of the questions which are being examined is described in general terms the first time it appears, but for all subsequent occasions only the facts pertinent to the question at hand are discussed.

1. Topic Effect on Essay Scores

Godshalk, Swineford and Coffman (1966) examined the influence of topic on essays written by grade eleven and twelve students who were native speakers of English. The students wrote six objective tests and five free-writing exercises. There were 25 readers who scored the essays independently and holistically. Six hundred and forty-six subjects wrote five writing exercises each and each exercise was scored by five different readers. The holistic scores based on total impression ranged from 1 for an "inferior paper" to 3 for a "superior paper". For one essay a subject's score could range from 5 to 15. For five essays his score might range from 25 to 75.

Two of the essays had a time limit of forty minutes and the other three essays had a time limit of twenty minutes. The different essays aimed at different kinds of writing, such as argumentation or description.

An ANOVA showed the scores for topic had a significant effect indicating that topic was a factor affecting scores. Therefore, the authors concluded that if different topics were used on different tests, a method of equating them would be required. The mean scores for the five topics were as follows: 9.41, 8.95, 8.93, 8.12, and 7.99.

There was no evidence that greater reliability was achieved with the two 40-minute essays (for a total of 80 minutes) than with the three 20-minute essays (for a total of 60 minutes).

Godshalk et al (1966) wrote:

> At the time the study was designed, it was known that the unreliability of essay tests came from two major sources: the differences in quality of student writing from one topic to another, and the differences among readers in what they consider the characteristics of good writing. The first source of error could be reduced by having students write on a number of different topics. Thus, an individual's rating would not depend on whether or not he could find something interesting and accurate to say on a single topic which he might never have considered before. (p. 5)

The authors also stated:

> The mean square for topics is significant,

> indicating that the ratings assigned varied from topic to topic. When the score is the sum of such ratings, one does not need to be concerned about such differences. The differences are significant, however, under certain other conditions. For example, if the five topics had been assigned as alternate topics from which one or two could be chosen by students, a student's rating might depend more on which topic he chose than on how well he wrote. Or if one topic has been assigned to one form of a test and another topic to a second form, then some method of equating the scores would be required; otherwise, the magnitude of an individual's score would depend partly on which form of the test he wrote. Differences among the difficulties of the topics would be part of the error. (p. 13)

The authors also commented on the usefulness of having students select their respective topics. They found:

> The significant student-by-topic interaction.... is interpreted as indicating that some of the students do relatively better on some topics while other students do relatively better on other topics. Could this problem not be overcome by offering the students alternative topics? Unfortunately not. In the first place, there is

> no evidence that the average student is able to judge which topic will give him the advantage. In the second place, the variability in topics already discussed would be introducing error at the same time that students might be eliminating error by choosing the topic on which they were most adequately prepared. (p. 13 & 14)

Tables 1 and 2 are reproduced from Godshalk et al. The first table gives the average correlations based on single readings for one topic and across topics. The second table gives the same correlations based on five readings. Godshalk et al. (1966) state:

> The diagonal entries in each table are estimates of reading reliability; the other entries are estimates of scoring reliabilities. Score reliabilities for single topics read once range from .221 to .308. Reading reliabilities for single topics read once range from .361 to .411. (p. 16)

Table 1

Average Correlation Based on Single Readings for One Topic.

Topic	A	B	C	D	E
A. Pen pal	.366				
B. Teen-ager	.237	.411			
C. Imagine	.242	.261	.385		
D. Step 2C	.253	.264	.260	.361	
E. Step 2D	.221	.308	.269	.297	.408
T. Total Essay Score	.518*	.574*	.554*	.561*	.581*

*Spuriously high because part is included in the total

Note. From Measurement of Writing Ability (p. 16) by Godshalk, F., Swineford, F., & Coffman, W., 1966, New York: College Entrance Examination Board.

Table 2

Correlations between Total Scores on the Five Readings of Each Essay Topic (Estimates of Reading Reliabilities are shown in Parenthesis).

Topic	A	B	C	D	E
A. Pen pal	(.743)				
B. Teen-ager	.466	(.777)			
C. Imagine	.483	.505	(.758)		
D. Step 2C	.516	.521	.523	(.739)	
E. Step 2D	.435	.584	.520	.592	(.775)
T. Total Essay Score	.738*	.791*	.777*	.804*	.921*

*Spuriously high because part is included in the total

Note. From Measurement of Writing Ability (p. 16) by Godshalk, F., Swineford, F., & Coffman, W., 1966, New York: College Entrance Examination Board.

Huddleston (1954) carried out a study involving 413 native speakers of English. Each student wrote three twenty minute essays. Each essay was on a different topic. On the first occasion each essay was scored once, but the essays were so distributed that the three essays of one subject were scored by three different readers. On a second occasion the essays were again scored so that each essay was scored twice. Two hundred and ninety-four subjects were involved in

this part of the study so that 1,764 essay scores were handed in. The intercorrelations below are based on scores given by readers across topics and so take both topic and reader into account as factors affecting essay scores. Huddleston (1954) writes

> ...the intercorrelations of the three essay questions are found to be .41, .41, and .32. If the Spearman-Brown prophecy formula is applied, and the assumption is made that an essay test has three twenty minute questions, each with a correlation of .41 with each of the other two, then the estimated reliability of the total test becomes .68. Such an estimate is the fairest available indicator of the true reliability of the essays used in this study. (p. 179)

In their survey of the literature, Ebel & Damrin (1960) also mention the effect of topic on reader reliability. They write, "The essay topic being marked when optional questions are provided from which the candidate is to select one, some topics are more reliably marked than others." (Ebel & Damrin, 1960, p. 1504)

Finlayson (1951) looked at the writing of 197 students who were native speakers of English and whose average age was twelve years, two months. The students wrote two essays one week apart. On each occasion they

were given a choice of four topics and on both occasions they were given one hour to complete the task.

Six experienced teachers scored the essays. Each teacher scored all the essays without any knowledge of how the other teachers were scoring. The teachers based their scores on general impression and used a scale ranging from 1 to 20.

An ANOVA was done using the factors, essay, student, and reader. Differences between readers were significant and the performance of students from essay to essay varied significantly. The essay X student interaction was significant at the .01 level. The author states, "From this finding we may conclude that the performance of a child on one essay is not representative of his ability to write essays in general... ."(Finlayson, 1951, p. 132)

Finlayson correlated the scores of the six readers individually for the two essays written by each subject. In Table 3 we find the reliability cofficients for each of six markers marking two essays per subject.

Table 3

Essay Reliability Coefficients for Each Marker.

Marker	M1	M2	M3	M4	M5	M6
r_{xy}	.724	.650	.673	.601	.798	.697

Note. From The Reliability of the Marking of Essays by Finlayson, D., 1951, British Journal of Educational Psychology, 21, p. 130.

Finlayson comments, "It will be observed that these correlations which take into account the fluctuations of the children from essay to essay are considerably lower that the mark re-mark correlations for one essay... . The mean correlation for the six markers was .691 as compared with .810." (Finlayson, 1951, p. 130)

In Finlayson's ANOVA, marker-essay interaction was also found to be significant. The author states that this shows that, "..the markers hold varying opinions amongst themselves as to the relative merits of each batch of essays."(Finlayson, 1951, p. 132)

Schonell(1942) conducted a study with native speakers whose ages ranged from 7 to 14. Among other tests, the 1300 students wrote on 4 topics which were

of 4 types: descriptive, narrative, imaginative and reproductive. The marking scheme was broken down into 12 marks for content, 7 for structure and 6 for mechanics. The author found that the number of words written varied from topic to topic. For 13 year olds (N=578) one topic averaged 301 words while another topic averaged 199 words. He also found that different topics yielded different types of errors. Furthermore, he found that the type of sentences written showed some variety for different topics while being, nevertheless, basically the same. Finally he ventured the opinion that a student's interest in a topic could influence his performance.

2. Occasion for the Reader (Intrareader Reliability)

In the study by Finlayson (1951) referred to earlier, the investigator had four of his readers score the same essays four months later to see whether or not time would affect their scoring. His results showed that the mean consistency was .810 and, as can be seen in Table 4, intrareader reliability for essay X ranged from .678 to .966 and for essay Y it ranged from .636 to .959.

Table 4

Mark Re-Mark Correlations.

MARKER	r_{xx}	r_{yy}
M2	.731	.766
M4	.678	.636
M5	.966	.959
M6	.887	.856

Note. From The Reliability of the Marking of Essays by Finlayson, D., 1951, British Journal of Educational Psychology, 21, p. 129.

In the summery referred to earlier, Ebel and Damrin (1960) mention Lambs's (1953) study and write, "When the same one-hour papers are remarked four months later by the same readers, the mean consistency of eight independent readers is .86; the mean consistency of four readers made up of all possible combinations of the eight independent readers is .96." (p. 1505)

3. Who the Reader Is

In the Godshalk et al. (1966) study, each student wrote five essays on five different topics. An ANOVA showed the variable "reader" to have a significant main effect. The interaction between reader and topic was also significant.

Interreader reliability for twenty-five readers with five essays per student and with each essay being scored by five different readers was estimated at .92. They stated, "This means that if a second group were chosen and the papers were read again, it might be expected that the two sets of scores would produce a correlation of approximately .921" (Godshalk et al., p. 12). At the conclusion of this study Godshalk et al., (1966) state:

> The reliability of essay scores is primarily a function of the number of different essays and the number of different readings included. If one can include as many as 5 topics and have each topic read by 5 different readers, the reading reliability of the total score may be aprroximately .92 and the score reliability .84 for these samples. In contrast, for one topic read by one reader, the corresponding figures are .40 and .25 respectively. The increases which can

be achieved by adding topics or readers are dramatically greater than those which can be achieved by lengthening the time per topic or developing special procedures for reading. (p. 39 & 40)

In the Finlayson study (1951), mentioned earlier, an ANOVA was carried out and as in Godshalk et al (1966), it was found that the variable reader had a significant main effect. The interaction between reader and essay was also found to be significant.

The author also pooled the scores of four readers on one essay and correlated them to the pooled scores of four other readers scoring the same essays. These correlations ranged from .712 to .937. The correlations for the same groups of readers reading two essays per subject ranged from .803 to .961.

Table 5

Means and Standard Deviations of Readers Scores Based on 394 Essays per Reader.

READER	MEANS		STANDARD DEVIATIONS	
	X essay	Y essay	X essay	Y essay
R1	7.558	7.766	3.174	2.817
R2	10.675	10.883	2.184	2.218
R3	10.609	10.310	3.475	3.262
R4	9.695	10.289	2.873	2.663
R5	9.340	9.594	3.149	2.991
R6	11.492	11.132	2.683	3.043

Note. From The Reliability of the Marking of Essays by Finlayson, D., 1951, British Journal of Educational Psychology, 21, p. 128.

Looking at Table 5 it can be seen that the readers were basically consistent in their scores across essays. Score magnitude across readers ranged from a low of 7.558 for reader one to a high of 11.492 for reader six when scoring essay X. The range for essay Y was 7.766 to 11.132.

For essay X, Finlayson also correlated the scores of six readers (N=197). The correlations ranged from .591 to .824. Finlayson (1951) states:

> The mean intercorrelation is .738, and this an estimate of the reliability of individual markers, for by comparing the marks of two random markers on the one test, we are, in fact, obtaining a measure of how the marks given to an essay by one marker compare with the marks given to the same essay by another marker." (p. 128)

In a large British Columbia study (1977), post-secondary students were asked to write a 300 to 500 word expository essay on one of five possible topics as part of their English Placement Test. These compositions were scored by 243 English Teachers. The readers were trained in the marking scheme using practice compositions. On the first weekend of scoring one reader marked for development and structure while a second reader marked the same compositons for sentences and words. The following weekend the scoring of the compositons was randomly redistributed with one reader scoring a composition for all four aspects.

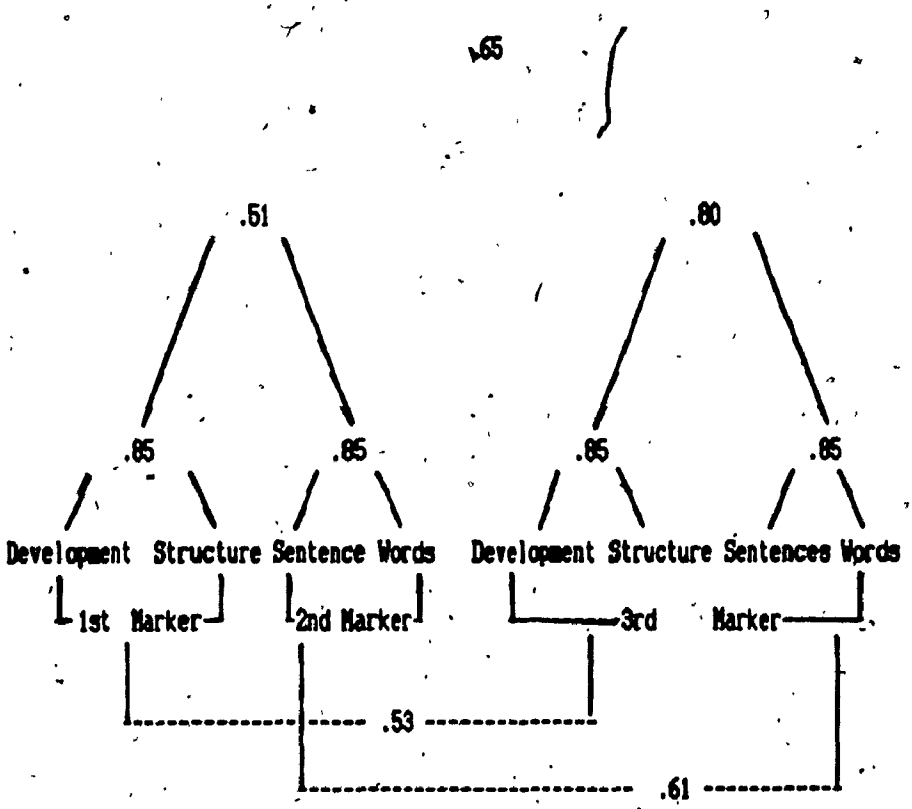
Each aspect had a discriptor allowing for a high score of 9 and low score of 1. Therefore, the maximum score for two readings was 72. The results shown in Table 6 indicate that an important variable in

determining the level of correlation in scores is who the marker is. When marker one marked for development and structure while marker two marked for sentences and words, their respective marks for the same essays had a 0.51 correlation. On the other hand when marker three marked for development and structure and for sentences and words too, the marks he gave had a 0.80 correlation. This 0.80 correlation of marks by the same marker but for different items of the essays is much higher than the correlations for different markers marking for the same items. With markers one and three marking the same essays for development and structure the correlation of their respective scores was 0.53. With markers two and three marking the same essays for sentences and words the correlation of their respective scores was a low 0.61.

In fact, all intrareader correlations whether or not they were based on the marking of the same items are 0.80 or higher, and all interreader correlations even when based on the marking of the same items were never higher than 0.65.

Table 6

English Placement Test - Correlation of Scores by the First, Second and Third Marker.



Note. From English Placment Test, 1977, Vancouver, Research Institute of British Columbia.

Interreader reliability coefficients for readers using the same scoring schemes are .53 and .61. Interreader reliability coefficients for readers using different scoring schemes are .51 and .65 when three readers are included and four scoring schemes are used.

In the Anderson and Follman (1967) study 10 essays were scored by five groups of readers, with each group using a different scoring scheme. Within each group there were five readers, with each reader scoring the essays independently. The average reliability scores were obtained for each group (scoring scheme) using Ebel's intraclass correlation method. In Table 7 we see the average reliability scores for one rater and five rates respectively for each scoring scheme. For four of the scoring schemes the reliability score results for five readers were very close, ranging from a low 0.930 to a high 0.953. Only the results from one scoring scheme stood out where for five readers the reliability score was 0.810.

Table 7

Essay Score Reliability for One and Five Readers Respectively for Five Scoring Schemes.

Scoring Scheme	1	2	3	4	5
One rater	.460	.727	.769	.788	.813
Five raters	.810	.930	.943	.949	.953

Note. From An Investigation of the Reliability of Five Procedures for Grading English Themes, Follman, J., C., & Anderson, J., A., 1967, Research in the Teaching of English, 1, # 2.

In another study, twenty essays were selected from those written by low-level ESL students and freshman remedial-English students who were native speakers. The essays were scored impressionistically by six readers using a scale from 1 to 20. (Brodkey and Young, 1981) The interreader correlations varied from a low of .49 to a high of .89 with a mean correlation of .73.

Perkins (1982) compared the holistic and analytic methods of scoring compositions. The subjects were 104 non-anglophone students enrolled in freshman composition courses at Southern Illinois University.

The scorers were graduate assistants. The holistic scheme was a five point scale with a brief descriptor for each point on the scale. Each compositon was scored by two readers independently from one another. The author states; "The interrater correlations ranged from .47 to .77 with a mean correlation of .58 yielding 34% true overlap between graders-- (.58) = .34" (Perkins, 1982, p. 9). Perkins indicated surprise and disappointment at the low interrater correlations.

In another study with native speakers of English, Diederich (1974), "secured 300 papers written by students in their first month at 3 different colleges and had them all graded by 60 distinguished readers in 6 occupational fields"(p. 5). Half the judges were academics and half were not. No discussion was held concerning grading standards. "Our only directions were to sort the papers into 9 piles in order of general merit. The only rules were that all 9 piles must be used, with not less than 12 papers in any pile." (p 5) Of the 300 essays, 101 received every grade from 1 to 9, 94% received at least 7 different grades, and no essay received less than 5 different grades. The median correlation of the readers' scores was .31. Diederich found that 43% of the variance was accounted for by the fact that different readers were being influenced by different factors such as sentence

structure, ideas expressed and vocabulary. This left 57% of variance unexplained.

Pike (1979) conducted a study which looked at the TOEFL Test with its sub-sections. As part of the study, the TOEFL Test was compared to other measures including essay tests. The subjects were 199 Japanese, 145 Chileans and 98 Peruvians.

At first Pike divided the task of scoring the essays into four distinctive parts: content-quality, content-quantity, form-quality and form-quantity. Finding that the headings of quality and quantity yielded virtually no distinction in scores, he reduced the parts to just form and content.

Each subject wrote four essays and each essay was read by eight readers; two for each part of the original scoring breakdown. In pooling the scores Pike found that the reliability for overall essay form was 0.98 for Peruvians, 0.92 for Chileans and 0.91 for Japanese. For content the reliabilities were 0.98, 0.96 and 0.91 respectively. These estimates are based on 16 readings for form per subject (4/essay) and on 16 readings for content per subject. (Japanese: N = 199, Chilean: N =145 and Peruvian: N = 98).

4. Scoring Schemes

Generally when compositions are scored, one of many possible types of scoring schemes is used in determining the quality of the writing. These scoring schemes act as guidelines directing the reader to what he should look for in the writing in order to give a score. The scoring scheme a reader uses influences how he scores an essay. For example, one scheme might emphasize the mechanical aspects of writing such as grammar while another scheme emphasizes style or content. A subject with strong mechanics and poor content would score well on the former scoring scheme but poorly on the latter scoring scheme. Hence, to some extent, the score a subject receives is in part determined by the scoring scheme used.

As was mentioned in chapter one, the purpose in scoring essays was to determine the quality of the writing. The scoring scheme used or developed should also reflect what kind of proficiency is being tested. We might be testing for subjects wishing to be admitted to an English language university, to become creative writers or maybe reporters.

Even if we agree that our purpose is to determine if a subject's writing is of a sufficient quality to work successfully at university standards, we might

still find ourselves in disagreement as to what a good essay is. Such differences would be reflected in the scoring schemes developed. Furthermore, even if we agreed on what constitutes good writing, we might disagree on how best to test and evaluate such writing.

Scoring schemes can be divided into two basic types. The first type is holistic. With such a scheme one or more readers score an essay based on a general impression. The scoring scheme acts as a guidepost directing this general impression. Kyle Perkins (1982) prefers holistic scoring schemes and feels that they have higher construct validity. Nevertheless, he states that, "Such evaluation can be highly subjective due to bias, fatigue, internal lack of consistency, previous knowledge of the student, and/or shifting standards from one paper to the next."(p 2,3)

The second basic type of scoring scheme consists of objective measures. These have the advantage of close to perfect reliability. One of the problems with such schemes is that their relationship to hand-writing ability is not obvious.

In this section, different ways of scoring essays are examined. Anderson and Follman (1967), whose study was mentioned earlier, had groups of readers score the same ten essays using five different scoring schemes. Each scheme was used by five readers.

Although different from one another the scoring schemes were all holistic. The correlations of rating group scores ranged from a low of .51 to a high of .99. In the case of four of the scoring schemes the correlations were all higher than .9. The fifth scheme accounted for all the lower correlations. (p 197)

Perkins (1982), whose study was also discussed earlier, had 102 non-native speakers write an essay which was scored holistically and analytically. The holistic scoring scheme was a five point scale with a brief descriptor for each point. The analytical scoring scheme had five categories with a range of scores of one to five for each. The categories were relevance, fluency, vocabulary, grammar and mechanics. Each category included a brief descriptor.

The subjects also wrote the TAS Test (Test of Ability to Subordinate) and their essays were examined for words per T-unit.

The data were submitted to regression analysis with the holistic scores as the dependent variable and all other scores as predictor variables. Analytical scores were found to be the single best predictors accounting for 89% of the variance in holistic scores. The best two variable predictor was analytical scores and vocabulary, together accounting for 93% of the variance in the holistic scores.

The author expressed surprise that two different scoring schemes basically assessed the same constructs.

In the British Columbia study (1977) mentioned earlier readers scored the same essays while using different criteria. The four criteria were development, structure, sentences and words.

Table 8

Correlation Coefficents for Scores Given by Readers on Essays Using Different Criteria.

	Readers	Criteria	Correlation
1.	1 to 1	development to structure	.85
2.	3 to 3	" " "	.85
3.	2 to 2	sentences to words	.85
4.	3 to 3	" " "	.85
5.	3 to 3	development & structure to sentence & words	.80
6.	1 to 2	" " " " " " "	.51
7.	2 to 3	sentences & words to sentences & words	.61
8.	1 to 3	development & structure to development & structure	.53

Note. From English Placment Test, 1977, Vancouver, Research Institute of British Columbia, p.52.

Correlation number six shows two different readers scoring the same essays using different criteria: result .51. In correlation numbers seven and eight we see different readers scoring the same essays with the same criteria: results .53 and .61. Correlations one through five are notably higher for the apparent reason that these correlations compare scores given by the same readers.

5. Comparing Objective and Composition Measures

Pike (1979) correlated essay scores with scores from the sub-sections of the TOEFL Test. These correlations were corrected for attenuation due to unreliability due to such factors as test length and range of a subject's scores on a measure.

Table 9

Observed Correlations Between Multiple-Choice Measures and Composition Measures. *

TEST		Corrected for attenuation	
		ESSAY	ESSAY
Multiple Choice		form / content	form / content
	English Structure	70 / 81	76 / 90
	Vocabulary	59 / 68	66 / 77
	Reading Comprehension	60 / 68	71 / 79

Note. From An Evaluation of Alternative Item Formats for Testing English as a Foreign Language, p. 66, by Pike, 1979, Princeton, N.J., Educational Testing Service.

*In the original table the correlation coefficients were shown separately for the Japanese, Chilean and Peruvian students. The correlations shown above are the composite measures for these three groups.

The data in Table 9 indicate that between 60 and 90% of the variance in the composition scores is accounted for by the variance in the sub-sections of the TOEFL Test.

The Michigan Test of English Language Proficiency Mnaual (MTELP Manual, 1977) shows intercorrelations for the sub-parts of the Michigan Battery including those for compositons and the MTELP.

Table 10

Correlation of MTELP Scores to Essay Scores with Means and Standard Deviations.

MTELP Form	Number of Subjects	Mean	SD	Intercorrelation with essay scores
F, G & H	1793	62.84	27.84	.908
D	3974	77.18	14.14	.696
E	3706	79.70	13.33	.681

Note. From Michigan Test of English Language Proficiency Manual, p. 11, 1977, English Language Institute: University of Michigan.

Huddleston (1954) also examined intercorrelations between essay scores and the Objective English Test. The correlation obtained was .60. This result is based on 294 subjects who wrote 3 essays.

The intercorrelation for essay 1 was .43; for essay 2 it was .48 and for essay three it was .47. The

correlation of .60 corrected for attenuation was .75. The Objective English Test dealt with such matters as punctuation, idiomatic expressions, grammar and sentence structures.

Chapter Three

The Study

In Chapter Two the designs and results of other studies were examined. In this chapter, the procedures used in this study are described. The chapter is divided into four headings: (a) the measures , (b) description of the scoring of the measures, (c) the statistical measures used to interpret the scores of the subjects on the measures, (d) a description of the applicants.

1. The Measures

The data for this study was obtained from a testing session at Concordia University where the applicants were being tested for their ability to meet

the language requirements for admission to the university. The test administered to the applicants was the Concordia English Language Diagnostic Test (CELDT). On the evening from which the data were obtained (March 28, 1980), the CELDT was composed of a PT and the MTELP. Ninety applicants completed the two tasks.

The MTELP has a multiple-choice format, and consists of one hundred items, each worth one point. For each item there are three distractors and one correct answer. Forty of the items fall under the heading of grammar, another forty fall under the heading of vocabulary and twenty fall under the heading of reading. The applicants are given seventy-five minutes to complete the task.

For the PT, the applicant is given a handout which instructs him to write a five hundred word composition on a given topic (see Appendix B for a sample of the handout given to the applicants). The topic includes a title, a short argument, and a counter-argument, and the applicant is instructed to select one or the other point of view. He is then expected to incorporate the ideas into his composition. He is given one hour to complete the task.

2. The Scoring

The MTELP answer sheets of the applicants are scored with the use of a template. The PT is scored independently by two readers who do not know the MTELP scores of the applicants. These readers are ESL teachers who have taken part in a three hour session during which the scoring scheme was explained (see Appendix C for a sample of the scoring scheme used by the readers). The scheme is holistic and is composed of a range of five possible scores with a brief descriptor for each point on the scale.

At a practice session, sample compositions were given to the readers who were then asked to score them with the scoring scheme as a guide. At first the scores tended to vary a lot from reader to reader but as the session proceeded the scores tended to become much closer to one another.

For this study, approximately half of the PTs were scored by two readers (A and B) with the other half being scored by two other readers (C and D). Six months after the PTs had originally been scored, readers A and B were asked to again score the same PTs they had scored six months earlier. Readers A and B were used for this score re-score part of the study

because they were available, while readers C and D were not.

The purpose in having the same two readers score the same PTs on two different occasions was to see whether or not the readers would significantly change the scores they gave. It was felt that an interval of six months would be sufficient to insure that the scores given on the first occasion would not be remembered and so influence the second set of scores. Also, given that these readers were regularly scoring other PTs over this six month period, it seemed unlikely that they would recall the individual PTs.

Half of the PTs scored by the readers on the second occasion were presented to the readers in the original script version. The other half were the same as the original script version except for the fact that they were typed (see Appendix A for the guidelines used to type the Pragmatic Tasks of the applicants).

The reason for having the readers score half of the PTs in a typed version and the other half in their original version was to determine whether or not legibility was a factor influencing scores on the PTs.

Forty-two PTs were scored on the second occasion by readers A and B. The PTs to be typed were randomly selected by taking the pile of forty-two PTs, shuffling them haphazardly on a table and then alternately

assigning them to one of two piles. The typed versions of the PTs were made from one of the two piles. This means that on this second occasion readers A and B read the same forty-two compositions. Both readers had read these compositions six months earlier. On this second occasion half of the forty-two compositions read and scored were typed.

All the scores obtained from the measures yielded the following data:

(a) the MTELP scores for 90 subjects

(b) the vocabulary sub-scores of the MTELP(n=70) (n=70, not ninety, because not all the applicants used the same form of the MTELP. In seventy cases the same form was used. It was felt that the scores of the sub-sections for different forms of the MTELP were not comparable as they were not equated. The scores of the MTELP as a whole are equated across forms.)

(c) the grammar sub-scores of the MTELP(n=70)

(d) the reading sub-scores of the MTELP(n=70)

(e) the scores on the PT scored by reader A on occasion one (n=46)

(f) the scores on the PT scored by reader B on occasion one (n=45)

(g) the scores on the PT scored by reader C on occasion one (n=45)

(h) the scores on the PT scored by reader D on occasion one (n=44)
(When the compositions were first read, readers A and B always read the same PTs with one exception. Reader A read one extra PT where the other reader was reader C.)

(i) the scores on the PT scored by reader A on occasion two with the format script (n=21)

(j) the scores on the PT scored by reader A on occasion two with the format type (n=21)

(k) the scores on the PT scored by reader B on occasion two with the format script (n=21)

(l) the scores on the PT scored by reader B on occasion two with the format type (n=21)

(m) the composite scores of readers A and B on the PT on occasion one (n=36) (Composite score means adding the scores of two

readers on the same PT and dividing the result by two)

(n) the composite scores of readers C and D on the PT on occasion one (n=34)

3. Statistical Measures

To interpret the scores the following statistical measures were used:

(a) correlations of scores given by one reader to the same PTs on two different occasions to examine intrareader reliability and to examine the question of format as a possible scoring factor.

(b) correlations of scores given by two readers to the same PTs to examine interreader reliability.

(c) means and standard deviations of PT scores to look at average score, magnitude, and the dispersion of the scores.

(d) an analysis of variance with reader and format (script/type) as independent variables with the PT scores as the dependent variable, to

look for main effects and possible interaction.

(e) correlations of PTs to MTELP sub-scores to see the extent to which scores on one test are predictable from the other test.

(f) multiple regressions with PT scores as the dependent variable and MTELP scores and sub-scores as the predictor variable to examine the question of unique information between the two tests.

(g) a Chi-square goodness of fit test was used to determine the scale quality of the data (See appendix E for a discussion on this matter.)

4.The Applicants

The applicants were students who were not native speakers of English who wished to be admitted to Concordia University. Approximately 42% of the applicants were native speakers of French. The others represented a cluster of twenty other languages. (see appendix D for a complete description of the native languages of the applicants.) The applicants had varied educational goals (see appendix F for more

details); 53% were female, 47% male; the average age of the group was 24.2 years of age.

Chapter Four

The Results

In this chapter the results of this study are shown. The results are (a) intrareader correlations, (b) interreader correlations, (c) a t-test between correlated pairs of means, (d) an analysis of variance, and (e) measures which deal with the question of independence.

1. Intrareader Correlations

Table 11

Intrareader Correlations for Readers A and B. (Product-moment Correlations with Spearman Rank-difference Correlations in Parenthesis)

Occasion One	Occasion Two					
	1.	2.	3.	4. (typed)	5.	6. (typed)
1. reader A (n=42)	.86 (.87)					
2. reader B (n=42)		.73 (.71)				
3. reader A (n=21)			.90 (.92)			
4. reader A (n=21)				.83 (.87)		
5. reader B (n=21)					.73 (.68)	
6. reader B (n=21)						.76 (.72)

The results in Table 11 are based on score re-score results. Readers A and B independently scored 42 PTs called Occasion One. Six months later the same two readers scored the same PTs a second time (Occasion Two). On Occasion One all the PTs were in the

original script. On occasion two half of the forty-two PTs were typed. Therefore reader A's score, re-score correlation for n=42 is based on half of the PTs being typed on Occasion Two. The same applies to reader B. Correlations number three and five correlate scores where on both occasions the PTs were not typed. Correlations where n=21 are sub-samples of the correlations where n=42.

2. Interreader Correlations

Table 12

Interreader Correlations for Readers Scoring the Same PTs. (Product-moment Correlations with Spearman Rank-difference Correlations in Parenthesis). Means and Unbiased Estimates of Population Standard Deviation(s).

			Reader		
			B (occasion 1) $\bar{X}$=2.86 s=1.20	B (occasion 2) $\bar{X}$=2.76 s=1.21	C $\bar{X}$=2.91 s=0.89
Reader	$\bar{X}$	s			
A(occ.1) (n=42)	3.19	1.24	.76 (.79)		
A(occ.2) (n=42)	3.24	1.08		.83 (.84)	
D (n=44)	2.98	1.04			.78 (.77)
All correlations are significant at the .001 level					

The results in Table 12 show that the scores of readers A and B correlated higher with one another when they scored the same 42 PTs six months after scoring them the first time. The average correlation for the four readers(parametric) is 0.79. For readers A and

B it is 0.795.[1] The average intrareader correlation for readers A and B is also 0.795. This indicates that at least for the two readers in this study average intra and interreader correlations are not different.

3. A t-test between Correlated Pairs of Means

Table 13

T-test Results for Differences Between Correlated Pairs of Means.

Reader		t-score	
A (occasion 1)	B (occasion 1)		
X = 3.19 S= 1.24	X = 2.86 S= 1.20	2.54*	(n=42)
A (occasion 2)	B (occasion 2)		
X = 3.24 S= 1.08	X = 2.98 S= 1.21	4.63**	(n=42)
C	D		
X = 2.91 S= 0.89	X = 2.98 S=1.04	4.56**	(n=44)

* significant at the .05 level
** significant at the .01 level

The results in Table 13 show that score magnitude is significantly affected by who the reader is. Hence even if the readers' scores were correlating

perfectly with one another , with these readers, score magnitude across readers would be a factor affecting an applicant's score.

4. An Analysis of Variance

A factorial analysis of variance with a two-way classification was performed to test if a difference in a reader or format had a significant effect on the score of the readers. Twenty-one applicants had their Pragmatic Tasks scored by readers A and B on Occasion One. On this occasion, all twenty-one Pragmatic Tasks were in the original hand-written version of the applicants. Six months later the same two readers scored the same twenty-one Pragmatic Tasks with the difference that on this occasion the twenty-one Pragmatic Tasks were typed. Otherwise there were no changes. Spelling and grammatical errors were left unchanged within the guidelines described in Appendix A. These results are shown in Table 14.

Table 14

Analysis of Variance for Reader and Format on Pragmatic Task.

Source of variation	sum of squares	degrees of freedom	mean square	F	significance of F
Main Effects	3.91	2	1.95	1.43	.25
Reader	3.86	1	3.86	2.83	.10
Format	.05	1	0.05	.04	.85
2-way interactions reader-format	.19	1	0.19	.14	.71
explained	4.10	3	1.37	1.00	.40
residual	109.14	80	1.36		
total	113.24	83	1.36		

The analysis of variance shows that neither reader nor format significantly affect the magnitude of the scores. Nevertheless, the F for reader is 0.10. Also, there is no significant interaction between reader and format.

5. Measures Which Deal with the Question of Independence

The results shown so far are concerned with possible sources of error in PT scores. The results in Tables 15, 16, 17 and 18 are concerned with the question of independence of the PT scores from the MTELP scores of the same applicants. These results are answers to these two questions: To what extent are Pt scores predicted by MTELP scores?, and To what extent do the PT scores give us unique information about the language abilities of the applicants not revealed by their MTELP scores?

Because of the results of the Chi-square goodness of fit test mentioned earlier, it seemed prudent to tabulate the results using both parametric and non-parametric measures. Table 15 appears to indicate that in nearly all cases the use of one measure or the other had little bearing on the results.

Table 15

Correlation of Reader Scores on the Pragmatic Task to MTELP Scores and MTELP Sub-scores: Grammar, Vocabulary and Reading. (Product-moment Correlations with Spearman Rank-difference Correlations are Below the Pearson Correlations)

Reader	A1 n=36	B1 n=36	C1 n=34	D1 n=34
MTELP	.65	.60	.67	.74
	.75	.63	.71	.76
MTELP sub-scores				
Grammar	.56	.42	.66	.75
	.59	.39	.65	.76
Vocabulary	.53	.39	.60	.64
	.58	.40	.60	.65
Reading	.61	.58	.57	.54
	.63	.54	.60	.52

Table 15 shows that the MTELP test as a whole has higher correlations with reader scores on the PT than any of the MTELP sub-scores. The Product-moment correlations with Spearman rank-difference for the MTELP test as a whole range from range from .63 to .76 with a mean of .71. Comparing this to interreader correlations of .79 and .77 (Table 12) it appears that

the readers correlate with one another just a bit better than they do with the MTELP scores.

Stepwise multiple regressions are reported in Tables 16, 17 and 18. In the procedure the computer statistical program selects the independent variable which best predicts the dependent variable and then displays the regression for this variable together with a constant. The program then selects the second best predictor, tabulates the beta weights and gives a multiple regression equation with the two independent variables and a constant. Finally the same is done including the third independent variable. At each point in the operation the level of significance is given for each beta weight. If the level of significance for any sub-score is less than .05 then it is left out of the multiple regression equation.

Table 16

Multiple Regression with the Composite PT Scores as the Dependent Variable and the MTELP Sub-scores as the Independent Variables. (n=70)

b_0 = constant = .93	Equation model:
b_1 = grammar = .05	$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$
b_2 = vocabulary = *	Equation with values:
b_3 = reading = .07	$Y = .93 + .05x_1 + .07x_2$
Y = dependent variable	$R^2 = .48$
X = independent variable	

* The vocabulary part of the MTELP is not included in the multiple regression equation because its level of significance was less than .05.

Table 17

Multiple Regression with the Composite PT Scores for Readers A and B only as the Dependent Variable and the MTELP Sub-scores as the Independent Variables. (n=36)

b_0 = constant = 1.47	Equation model:
b_1 = grammar*	$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$
b_2 = vocabulary*	Equation with values:
b_3 = reading = .14	$Y = 1.47 + .14x$
Y = dependent variable	$R^2 = .38$
X = independent variable	

* The vocabulary and grammar parts of the MTELP are not included in the multiple regression equation because their respective levels of significance were less than .05.

Table 18

Multiple Regression with the Composite PT Scores for Readers C and D only as the Dependent Variable and the MTELP Sub-scores as the Independent Variables. (n=34)

b_0 = constant = .89	Equation model:
	$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$
b_1 = grammar = .09	
b_2 = vocabulary*	Equation with values:
	$Y = .89 + .09x$
b_3 = reading*	
Y = dependent variable	$R^2 = .59$
X = independent variable	

* The vocabulary and reading parts of the MTELP are not included in the multiple regression equation because their respective levels of significance were less than .05.

Table 16 shows that, in taking into account the four readers, 48 percent of the variance in PT scores is accounted for by an optimum weighing of the values of the sub-scores of the MTELP. This means that the remaining 52 percent of variance must account for the unique information in the PT scores minus the variance due to error.

Table 17 shows that for readers A and B only 38 percent of the variance in the PT scores is accounted

for by MTELP scores. But with readers C and D (Table 18) 59 percent of the variance is accounted for by the MTELP scores.

6. Summary of the Chapter

In this chapter the results of this study were shown. In the next chapter these results are discussed and compared with the results of other studies reviewed in Chapter Two.

Chapter Five

Discussion and Conclusion

In this chapter, the results of this study are examined. Also, the results of other studies, reviewed in chapter two are looked at. Taking both into account, conclusions are drawn concerning the amount of variance in PT scores due to scoring error and the amount of variance in the PT scores which can be predicted by multiple-choice scores.

The variance due to topic effect, intrareader reliability and interreader reliability is examined. Next, the amount of variance in PT scores which can be predicted by the MTELP sub-scores is examined, and finally both are brought together to determine the percentage of reliable variance in PT scores which is due to an applicant's writing ability and independent of his score on the MTELP.

1. Topic Effect on Pragmatic Task Scores

An applicant's score on a PT is ideally an accurate reflection of that applicant's ability to employ the English language for university reading and writing. In the human sciences, complete validity and reliability in testing can never be attained. Therefore it is important to isolate undesirable factors and to determine the magnitude of their effects.

An applicant's score on the PT is determined on the basis of his writing on one topic only. The question then arises of whether or not his score might have been different if he had written on a different topic. The data from this study does not permit us to examine topic effect on PT scores. Fortunately, other studies have examined this question. In the articles mentioned in the review of the literature there is a consensus that the factor topic does affect scores. Of those which did look at topic effect, the two which I found the most informative were the Godshalk et al. (1966) and Finlayson (1951) studies.

In the Godshalk et al. study the applicant's wrote on 5 topics while in the Finlayson study each applicant wrote on 2 topics. The two studies used analysis of variance to examine their respective results. In both

cases there was significant interaction between topic and reader. This means that a given topic affects readers in different ways. On one topic, reader X might tend to give higher or lower scores to one set of essays, with reader Y giving higher or lower scores to different essays. Hence some of the variance in an applicant's score would be accounted for by the way a given topic influenced a reader's judgement.

Subject interaction with topic was also significant in both studies, indicating that some subjects did relatively better on one topic while other subjects did better on other topics. Therefore, we are not merely dealing with one topic uniformly having the effect of lowering or raising scores but of subjects being affected differently by different topics.

In addition to significant interaction, Godshalk et al. (1966) found a significant main effect for topic. Finlayson (1951) did not. If topic were seen to have a main effect it would mean that the magnitude of the scores would have a tendency to be higher or lower depending on which topic was used in testing. Considering that the subjects in the Godshalk et al. (1966) study wrote on 5 topics and that the subjects in the Finlayson (1951) study wrote on only 2 topics it seems quite possible the Godshalk study is a better indicator.

Furthermore the subjects in the Godshalk et al. (1966) study were in their final year of pre-university studies while the subjects in the Finlayson (1951) study had a mean age of approximately twelve years. Hence Godshalk's subjects more closely resemble the subjects of this study.

Also, Finlayson's subjects, who wrote two essays each, did not all write on the same topics as on both occasions they had 5 topics to choose from. The five topics on each occasion are quite different from one another but resemble closely the topics available on the second occasion. From the first set a subject could have selected the title 'Our Christmas Party' and on the second occasion he could have selected the similar topic 'Our Picnic'. Another example of two such topics are,'An old school bag - tell the story of its life' and 'A Christmas tree - tell the story of its life'.

In such a case we are comparing scores where some students have written on similar topics while others have not, and where readers are scoring subjects on different topics where they, the readers, might be influenced by the topic chosen when they score the essay. Furthermore, some subjects might be choosing topics they do well in while others might not be good at choosing topics.

These factors confuse the issue and are not indicative of the situation faced by aspiring students at Concordia University where topic choice is not part of the testing format.

Finally the Godshalk et al. (1966) study is based on more topics with more readers scoring each essay so that its results are probably more reliable.

The average intrareader reliability in the Finlayson study (1951) was 0.810 (Table 4, p. 28). In this study the average intrareader reliability was 0.795 (Table 11, p. 57). These results are close to one another.

Finlayson (1951) also compared the scores each reader gave across topics and found that the average correlation fell from 0.810 to 0.691 for a difference of 0.119. Seeing that the intrareader reliability results in the Finlayson study are close to those in this study, my best guess would be that the amount of error variance due to topic is something like that found in the Finlayson study: that is 0.119.

In conclusion, I would therefore say that, based on both Finlayson (1951) and Godshalk et al. (1966), topic interacts with both readers and subjects and, based on Godshalk et al. (1966), I would also say that topic has a significant main effect.

2. Intrareader Reliability

Reliability coefficients reflect only the ordering of two or more sets of scores. Hence, a reader's two sets of scores for the same PTs could theoretically be a perfect 1.00 with the mean score for the two sets being significantly different. Score reliability only compares rankings. It makes no reference to the comparative magnitude of the scores.

Still, differences in means for scores are relatively easily dealt with so long as a reader is consistently high or low in his scores. Adding or subtracting from his scores with a constant to bring them in line with the scores of other readers is not too difficult. Poor reliability presents a much more complex problem, and cannot be so easily corrected.

In this study readers A and B scored the same PTs on two occasions approximately six months apart. Given that these readers were regularly scoring other PTs , it was felt that this six month interval was sufficient to prevent them from recalling the scores they initially gave the PTs.

On the first occasion all 42 PTs were scored in the script format. On occasion two half of them were scored in the type format with the other half again being scored in the script format. But since format

has not proven to affect scores predictably and significantly, the change of format on occasion two can be ignored.

Reader A's correlation for the scores he gave the same 42 PTs, six months apart, was 0.86. For the same PTs, reader B's correlation was 0.73. Averaging these two results we obtain a mean correlation of 0.795, which is the best guess this study allows us to make about intrareader reliability.

Finlayson (1951) had four of his readers score the same essays four months apart. As the subjects had written on two topics, correlations were tabulated separately for each topic. With the first topic, intrareader correlations ranged from 0.68 to 0.97 with a mean of 0.81. For the second essay, the range was from 0.64 to 0.96 with a mean of 0.80.

These two mean correlations (0.80 & 0.81) are virtually the same as those obtained in this study (0.795).

Ebel and Damrin refer to a study where the same essays were scored four months apart. The mean consistency for eight independent readers was 0.86.

Standing alone, the intrareader reliability results from this study are limited in scope, as they are based on two readers only. But, when compared to the results obtained in these two other studies, they

appear to be representative of the results one might expect.

Still, there is a notable difference between readers A and B in their respective intrareader correlations (0.86 & 0.73). A further study with more readers would be of interest in determining whether or not variability in intrareader reliability was a significant factor affecting PT scores.

3. Who the Reader Is

In the Finlayson (1951) study, an analysis of variance was performed, and reader was found to have a significant main effect. The Godshalk et al. (1966) study also included an analysis of variance and it, too, found a significant main effect for reader. In this study an analysis of variance did not show reader to have a significant main effect.

Finlayson's (1951) study was based on 394 essays being scored independently by six readers for a total of 2364 scores. In the Godshalk et al. (1966) study the results were based on 646 essays being scored independently by twenty-five readers for a total of 16150 scores.

Considering these results, it is surprising that for the analysis of variance in this study the factor

'reader' does not have a significant main effect. Nevertheless, the t-test of this study does show reader to have a significant main effect. Hence, it is reasonable to assume that a 3-way repeated measures design (with writer as a factor) should give a significant F for reader. The t-test from this study is more sensitive than the analysis of variance and so it is concluded that, like Godshalk and Finlayson, reader in this study does have a significant effect on the scores of the PTs.

Table 19

Comparison of Intrareader and Interreader Correlations for This Study and Others Examined in Chapter Two

THE STUDY	CORRELATIONS		
	Intrareader	Interreader	Difference
This study	0.795	0.790	0.005
Finlayson(1951)	0.810	0.738	0.072
English Placement Test (1977)	0.810	0.510	0.300
Brodkey & Young (1981)	-	0.730	-
Perkins (1982)	-	0.580	-
DIEDERICH (1974)	-	0.310	-

In Finlayson (1951) and in the English Placement Test (1977) it has a notable effect and in the other

three studies noted, the interreader correlations are lower than those of this study. Diederich's results are probably accounted for in large by the nature of his study and in Perkin's (1982) case the author was disappointed with the low correlations of his readers.

4. The Idependence of Composition Scores to Multiple-Choice Test Scores

Pike (1979) correlated scores on essays to scores by the same subjects on a multiple-choice test which had three sub-sections (Table 9). The sub-heading titled English Structure had the highest over all correlation which was 0.755 and the sub-heading titled Vocabulary had the lowest correlation, which was 0.635. The correlations in this study between the scores on the sub-sections of the MTELP and the scores on the PTs had the same ordering. The sub-heading titled Grammar had an average correlation of 0.60 with the PT scores of the four readers. The sub-heading titled Vocabulary had an average correlation 0.54.

The average correlaion of the MTELP across forms (Table 10), (not from this study) to essay scores was 0.76. In this study the correlation of the MTELP to the scores of the four readers on the PTs was 0.67.

These results indicate that the scores of the readers in this study, like the readers in Pike's study, (1979) correlate best with what is titled structure or grammar on multiple-choice tests, and least with what is titled vocabulary. Furthermore, although lower, the correlations of the MTELP scores to the PT scores in this study are similar to those found in the MTELP Manual (1977).

The lower average correlations between the MTELP and the PTs in this study might be due to one of, or a combination of, several factors. The difference may be due to the fact that the correlations from this study are based on only four readers. Maybe correlations based on a larger pool of readers would yield different results. Another possibility is that the PT is sufficiently different from the standard essay so that its results are based on factors sufficiently different so as to correlate lower to the MTELP. A final possibility is that the PT scores in this study had less reliable variance than the essays correlated to the MTELP. If this were the case, then there would be less reliable variance for the MTELP to correlate too and hence the lower correlations.

Still, the average interreader correlation for this study (Table 12) is 0.79 and the average correlation of the scores of the four readers on the PT

to the scores on the MTELP is 0.67. Hence, the difference between how readers correlate amongst themselves to how they correlate to MTELP scores is only 0.12 (comparing the scores with Spearman rank-difference correlations the difference is only 0.09). It is interesting that when the readers are scoring the same PTs, their respective scores correlate only 0.12 higher than when their scores are correlated to scores on a completely different test. The 0.79 correlation between the two readers is the best guess of the reliability of PT scoring-taking format, occasion for the reader, and who the reader is into account. The multiple regression in table 16 shows that 0.48 of the variance in the PT scores is accounted for by an optimum weighting of the MTELP sub-scores. This represents common variance. But common variance can only be reliable variance and therefore 0.48 represents not 48% of this reliable variance but 0.48 divided by 0.79 which equals 0.61. That is, in this study, 61% of the reliable variance of the the PT scores and the MTELP is shared. Furthermore, 0.79 is our best guess of the reliable variance for the readers in this study without taking into account possible error due to topic. In the Finlayson study interreader correlations dropped an average of 0.12 (0.81 - 0.69) when topic was taken into

account. Interreader correlations in the Finlayson study were nearly identical to those in this study and so 0.12 is our best guess of the amount of error that might be expected if this study had permitted us to examine interreader correlations across topics. This means that 0.79 is probably a high estimate of PT score reliability and that if topic were considered as a factor affecting scores then approximately 0.67 (0.79 - 0.12) would be a better estimate of score reliability. Hence 61% shared reliable variance between PT scores and MTELP scores is extremely conservative. For of the remaining 39% of variance, some is unreliable due to the effect of topic.

This study indicates that 61% of the reliable variance in the PT scores can be predicted by the MTELP sub-scores. Therefore, with an optimum weighting, approximately 60% of whatever is being measured by scores on the Pragmatic Task can be accounted for by whatever is being measured by the MTELP.

Looking at Tables 16, 17 and 18 we see that in all three cases the vocabulary sub-score of the MTELP does not figure as a reliable predictor of the PT scores. Therefore, at least for the four readers in this study, it appears that whatever they are looking for when scoring the PTs is not associated with the skills that

determine success in obtaining a good score on the vocabulary sub-section of the MTELP.

Furthermore, the overall best single predictor (Table 16) of PT scores is the grammar sub-section of the MTELP. Nevertheless, just looking at the two pairs of readers in this study there are some notable differences. With readers A and B, the best overall predictor is whatever is measured by the reading sub-section of the MTELP. With readers C and D, the best overall predictor is whatever is measured by the grammar sub-section of the MTELP. This seems to indicate that different readers are looking for different things in determining the score they are going to give to a PT.

Also, the R^2 for all the composite scores on the PTs is 0.48 and on the basis of this study this is our best guess of the amount of variance in the PTs which can be predicted by MTELP scores. Still, results vary between readers. With the scores of readers A and B as the dependent variable, the R^2 is only 0.38. But with the PT scores of readers C and D as the dependent variable the R^2 is 0.59.

Since the MTELP sub-sections were the same for all three multiple regressions, these differences in R^2 must indicate that, despite the scoring guide the readers use to interpret the PT scale, who the reader

is plays a role in determining what factors influence a score on a PT.

5. Accounting for the Variance in Pragmatic Task Scores

A multiple regression was done with the composite reader scores as the dependent variable and with the MTELP sub-scores as the independent variables. The R^2 was 0.48. This means that 48% of the variance found in the composite reader scores can be accounted for by an optimum weighting of the MTELP sub-scores.

For 21 of the applicants it is possible to examine the correlations-taking format (script or type) and who the reader is into account. In Table 12 we see that the correlation for reader A on Occasion One and reader B on Occasion Two is 0.83. In the same Table we see that the correlation for reader A on Occasion Two and reader B on Occasion One is 0.81. Therefore our best guess as to the reliability of one reader, on one occasion, and taking format into consideration, is the mean of 0.83 and 0.81: that is 0.82. This means that the best estimate of the reliability of any one reader for this study is 0.82. It also means that, at least in this study, 0.18 of a reader's variance is unreliable and hence a source of error. The possible

sources of error taken into account are who the reader is, format and occasion for the reader.

In fact, though, an applicant's PT score is determined by two readers, not one. Statistically, reliability should increase when we have more readers. Hence, the composite scores of two readers should yield more reliable scores than any one reader. To determine the reliability of two readers we use the Spearman-Brown Prophecy Formula.

$$\text{Reliability of two readers(composite)} = \frac{2 \times 0.82}{1.82} = 0.90$$

Thus the reliable variance for two readers is 0.90. Of this variance, 0.48 is accounted for by an optimum weighting of the MTELP sub-scores. This leaves 0.42 (0.90 - 0.48) of the variance in reader scores not accounted for by the MTELP sub-scores and possible error due to format, occasion for the reader and who the reader is.

The question now arises how much of this variance (0.42) can be attributed to the skills of the applicants on the PT and how much might be attributed to other possible sources of error not examined in this study.

6. Conclusions

1. In this study it was found that the average intrareader reliability for two readers is 0.795. Still, one reader's intrareader correlation was 0.73 while the that for the other reader was 0.86. Also the mean intrareader correlation in the Finlayson (1951) study was 0.81, with a range from 0.73 to 0.86. Therefore, I think that it would be useful to examine further this possible source of error to determine whether or not some readers might be significantly less reliable than other readers.

2. With two readers and a small number of PT scores, legibility did not prove to be a factor reliably affecting how PTs were scored.

3. A t-test (Table 13) showed that for the four readers in this study score magnitude is significantly affected by who the reader is. Therefore, it might be advisable to use a constant to equate reader score values to insure that the magnitude of a subject's score is not based on who the reader is.

4. Based on the four readers examined in this study our best guess is that 48% of the total variance in PT scores can be predicted by an optumim weighting of the sub-scores in the MTELP. This is lower than what was

found in other studies. (See MTELP Manual, 1977, Pike, 1979, Huddleston, 1954, Perkins, 1982). With most of these studies this might be accounted for by the fact that the objective measures used were quite different than the MTELP which was used in this study. But the correlation of MTELP scores (Table 10) to essay scores (not this study) ranged from 0.681 to 0.908, which is notably higher than the results found in this study. This might be reflecting the fact that the PT is tapping different skills than the usual essay-type question does.

5. Sixty-one percent of the reliable variance in PT scores in this study was accounted for by an optimum weighting of MTELP sub-scores. This leaves 39% of the variance in PT scores as possibly yielding unique information concerning the English language writing abilities of the subjects.

But Godshalk et al. (1966) and Finlayson (1951) found error variance in essay scores due to changes in essay topics. It therefore seems likely that some of the remaining 39% of variance in PT scores is error due to topic and topic interaction with both readers and subjects.

6. About 61% of the reliable variance in PT scores is predicted by, or in common with, the variance in MTELP

scores. If one supposed in the area of about 0.12 in the lowering of the reliability of PT scores when topic for reader was taken into account, then the common reliable variance between PT scores and MTELP scores would be in the order of 72% (0.48/ 0.79 - 0.12). This means that in considering error variance for occasion for the reader, who the reader is and format in this study, and topic effect in other studies, our best guess is that 72% of the reliable variance in PT scores is shared with the variance in MTELP scores.

Hence, it can be concluded that the scores on the Pragmatic Task do yield information which is reliable and unique to the information obtained from the MTELP scores. In fact, considering our best guess as to the reliable variance in PT scores it can be said that somewhere in the range of 28% of this reliable variance is unique to the variance on the MTELP scores. Therefore, the PT scores do yield both unique and reliable information concerning the English Language abilities of the applicants.

The purpose of this study, as stated on page 3 , is to investigate the extent to which PT scores are independent of MTELP scores and the extent to which this independence is attributable to factors other than writing ability in English. These two principle questions were examined and the results were discussed

In this chapter. It was found that approximately 67% of the variance in PT scores is reliable and that 28% of this reliable variance is idependent to scores on the MTELP.

References

Allen, J. P. B., & Davies, A. (Eds.). (1977). The Edinburgh Course in Applied Linguistics: Testing and Experimental Methods (Vol. 4). London: Oxford University Press.

Brodkey, D., & Young, R., (1981). Composition Correctness Scores. TESOL Quarterly, 15, 159-167.

Diederich, P. B., (1974). Measuring Growth in English. National Council of Teachers of English. Urbana, Illinois.

Ebel, R. L., & Damrin, D. E., (1960). Tests and Examinations. Encyclopedia of Educational Research. Harris, 1502-1515.

Educational Research Institute of British Columbia. (1977). English Placement Test: Project Technical Report for May Test. Vancouver: Author.

Finlayson, D. S. (1951). The Reliability of the Marking of Essays. British Journal of Educational Psychology. 21, 126-134.

Follman, J. C., & Anderson, J. A. (1951). An Investigation of the Reliability of Five Procedures for Grading English Themes. Research in the Teaching of English. 1. No. 2, 190-200.

Godshalk, F., Swineford, F., & Coffman, W. (1966). Measurement of Writing Ability: College Entrance Examination Board. New York.

Huddleston, E. (1954). Measurement of Writing Ability at College Entrance Level: objective vs. subjective testing techniques. Journal of Experimental Education. 22, 87-98.

Kerlinger, F. N. (1964). Foundations of Behavioral Research: Educational and Psychological Enquiry. New York : Holt, Rinehart and Winston, Inc.

Lamb, H., (1953). The English Essay in Secondary Selection Examinations: A comparison of Two Methods of Marking. British Journal of Educational Psychology, 23, 131-33.

Michigan Test of English Language Proficiency Manual. (1977). English Language Institute, University of Michigan.

Perkins, K., (1982). An Analysis of the Robustness of Composition Scoring Schemes. Paper presented at the Annual Convention of Teachers of English to Speakers of Other Languages.

Perkins, K., (1983). On the Use of Composition Scoring Techniques, Objective Measures, and Objective Tests to Evaluate ESL Writing Ability. TESOL Quarterly. 17, 651-671.

Pike, L., W., (1979). An Evaluation of Alternative Item Formats for Testing English as a Foreign Language. Educational Testing Service. Princeton, N.J., TOEFL-RR-2.

Schonell, F., J. (1942). Backwardness in the Basic Subjects. London: Oliver & Boyd.

Appendix A

The Typed Pragmatic Tasks

(Steps taken in typing the Pragmatic Tasks)

a) The Pragmatic Tasks scored by readers A and B were chosen because these two readers were available six months later to re-score the Pragmatic Tasks. The forty-two PTs they had originally scored were randomly divided into two piles. One of these piles became the basis for the typed PTs.

b) The same number of words written on a line by an applicant were typed on a line.

c) If a line was skipped in the original version a line was also skipped in the typed version. But, if in the original version the applicant skipped every second line, no lines were skipped in the typed version. Also, if the applicant skipped more than one line only one line was skipped in the typed versions.

d) All errors were copied faithfully. For example, all syntactic, lexical and orthographic errors were typed as they occurred.

e) If an applicant failed to write a capital or put in a period at the end of a sentence the same was done in the typed version.

f) In one PT an applicant used the mathamatical symbol for therefore. In the typed version therefore was typed as follows: *therefore*.

g) If a word could not be deciphered it was replaced with the symbol > to indicate that a word was supposed to be there.

h) The above was explained to the readers.

Appendix B

Handout Given to the Applicants at the Testing Session Giving the Title of the Pragmatic Task and the Instructions on How to Write It.

Read the following paragraphs which present an argument and a counter-argument for the same topic. Select one point of view and write an essay of not less than 500 words which supports the point of view which you have adopted.

Your essay will be judged on:

1. accurate use of English language
2. organization of your ideas to present a clear and logical argument
3. the way in which you incorporate the ideas of the authors of the two passages so that they strengthen and add interest to your own presentation.

TOPIC: NO ONE WANTS TO LIVE TO BE A HUNDRED

A. Argument

For most of us, the prospect of growing old is horrifying. Who wants to live long enough to become incapacitated and senile? Is it really a good thing to extend human suffering, to prolong life, not in order to provide joy and happiness, but to give pain and sorrow?

Consider an extreme example of a person who is so senile that he has lost all his faculties, who exists merely in an unconscious state, and yet who is kept alive by artificial means for an indefinite period. Although friends, relatives, and even doctors agree that death will bring him release, everything is done to perpetuate what has become a meaningless existence. Might it not be preferable to let nature take its course in such cases, where death will relieve suffering?

B. Counter-Argument

Arguments for euthanasia, or "mercy-killing", of the elderly and infirm go against man's fundamental desire to continue living. Our natural tendency in any circumstance of survival is to cling to possibility for life. The right to live is basic to all humans, and it should be applied equally to the aged as well as

the young. Would you allow a sick baby to die because it is suffering? If not, why do so with a person who is elderly? We simply cannot make a different set of rules for the young and the old, nor do we have the right to make decisions about the lives of others. The duty of modern medicine is to prolong life, and each individual deserves the privilege to live out his or her life to the fullest.

Appendix C

Scoring Guide Employed by the Readers

Rating	Description
	Competent Writer. Develops a logical thesis and argument systematically with well-structured main and subordinate themes and relevant supporting detail. Accurate and appropriate language, lexis, grammatical patterns, layout and style. The occasional slip or infelicity reveals he is not a native writer. Often approaches bilingual competence. Responds well to tone and purpose of writing task.
	Modest Communicator in Writing. Conveys basic information competently, but logical structure of presentation will lack clarity. Work will show several slips or formal errors. Obvious limitation of style and lack of mastery of appropriate idiom in an otherwise intelligible presentation. Responds to tone and purpose of writing task but without consistency. Essay may well lack in interest and refinement of expression but

the basic message gets through.

Marginal Writer. Presentation has coherent appearance and several factual statements can be sequentially made. Work lacks logical structure. Frequent lexical and grammatical errors. Uses basic puctuation conventions but inaccurately. Will backtrack and repeat. Basic theme is conveyed but imperfectly.

Extremely Limited Writer. Produces a string of sentences rather that an essay. Some theme but not logically presented. Use of simple sentence structure and restricted lexis with errors and inappropriacies abounding. Demonstrates considerable difficulty in the conveyance of straightforward information.

Non-Communicator. Not able to write. Little evidence of having understood instructions for task, or response is virtually unintelligable. Length may fall exceedingly short of that required.

Appendix D

Information about the Applicants Who Wrote the PT and the MTELP

A. Their Purpose: The applicants are people whose native language is not English and who applied to be admitted to Concordia University. On the basis of the results of the CELDT battery the TESL Centre makes a recommendation to the Admissions office concerning the placement of the applicants.

B. The Number of Applicants: The sample is drawn from the testing session of March 28, 1980. Ninety-six people showed up to be tested that evening. Ninety of these applicants listed their native language as other than English and completed the PT and the MTELP. These ninety are the applicants used for this study.

C. Native Language:

The applicants who took part in this study represent twenty-one different languages. These are listed in table 20 along with the number of speakers representing each language.

Table 20

Native Languages of the Applicants Used for this Study

Native Language	Number of Applicants
French	38
Arabic	8
Spanish	7
Vietnamese	7
Iranian	4
Italian	4
Greek	3
Armenian	2
Chinese	2
German	2
Indonesian	2
Bengali	1
Czech	1
Dutch	1
Hebrew	1
Polish	1
Russian	1
Tagalog	1
Telugu	1
Turkish	1
Urdu	1

D. Choice of Degree: Before writing the CELDT battery the applicants are asked to fill a short questionaire. One of the sections asks the question 'Degree sought'.

The applicants have eight possible choices. Their choices are shown in Table 21.

Table 21

The Number of Applicant Responses for Each Catagory of the Question 'Degree Sought'

Catagory	Number of Responses
Administration	5
Arts	23
Commerce	11
Computer Science	9
Education	8
Engineering	12
Fine Arts	18
Science	4

E. Sex: The ninety subjects are composed of forty-eight females and forty-two males.

F. Age: The average age of the applicants is 24.2 years old.

Appendix E

Discussion on the Use of Parametric and Non-Parametric Statistics

Parametric statistics make more assumptions about the data of a study than non-parametric tests do. "In general the more assumptions a test makes the more powerful it is. In cases where both parametric tests and non-parametric tests are applicable, parametric tests are more powerful than non-parametric tests." (Allen & Davies, 1977, p. 184)

To be able to use parametric tests one must have interval data. The data of this study is found in two scales. One sale is that of the MTELP with possible scores ranging between 0 and 100. But this must be qualified. The MTELP is a multiple-choice test and for each question the applicant is given four possible answers to choose from. Furthermore the MTELP has one hundred questions with each question having a value of one. If a person were to make a guess for each question on the answer sheet without looking at the questions or without any knowledge of English whatsoever his most probable score would be twenty-five. Such a score would be uninterpretable. A

person scoring less than twenty-five might be unlucky or he might possibly be affected by the distractors.

The second scale for this study is that of the Pragmatic Task. Here the scores range from one to five with one being the lowest score and five the highest. Do these two scales represent interval data?

The simplest kind of scale is that for nominal data. "Examples of nominal categories are nationality, school attended, of age/not of age, pass/fail."(Allen & Davies, 1977, p. 118) In such a case the categories must be distinct only. One does not say that one nationality is higher or lower than another as one might say that a score of five on the Pragmatic Task is higher than a score of one. Clearly, in this study we are not dealing with a nominal scale which would not permit the use of parametric tests.

"Ordinal scales have the same properties as nominal scales plus the property that the values have an ordered relationship to one another."(Allen & Davies, 1977, p. 119) This is certainly true for the two scales in this study. They both have an ordered relationship in that on the Pragmatic Task a five is better than a four which in turn is better than a three and on the MTELP a ninety-five is better than a ninety and a sixty-nine is better than a sixty-eight. But an

ordinal scale is insufficient to permit the use of parametric tests.

One of the assumptions for parametric tests is that. "the samples with which we work have been drawn from populations that are normally distributed." (Kerlinger, 1964, p.258) But a bit further the author writes, "The evidence to date is that the importance of normality and homogeneity is overrated, a view that is shared by the author." In backing this statement the author cites three studies by Boneau dated 1953, 1960 and 1961.

"Interval or equal-interval scales possess the characteristics of nominal and ordinal scales, especially the rank-order characteristic. In addition, numerically equal distances on interval scales represent equal distances in the property being measured." (Kerlinger, 1964, p. 424) For the PT the distance between a four and a five is numerically equal to the distance between a two and a one and in the MTELP the distance between ninety-nine and ninety-eight is numerically equal to the distance between fifty-five and fifty-four, but do they represent equal distances for the properties being measured?

Some statisticians believe that all educational and psychological data can be ordinal at best. But Kerlinger writes, "Mostly nominal and ordinal are

used, though the probability is good that many scales and tests used in psychological and educational measurement approximate interval measurement well enough for practical purposes ." (Kerlinger, 1964, p. 425) "The lack of equal intervals is more serious since the distances within a scale theoretically cannot be added without interval equality. Yet, though most psychological scales are basically ordinal, we can with considerable assurance often assume an equality of interval." (Kerlinger, 1964, p. 426)

Hence strictly speaking the data of this study probably do not constitute an interval scale, nevertheless, if interval scale is not assumed we are forced to abandon some very useful statistical tests. It is the opinion of Kerlinger that most researchers presume interval data when in fact they have ordinal data. In cases where there are not gross differences between intervals he does not feel that it is a problem. In many studies one finds researchers using ANOVAs, Pearson Correlations and other parametric tests when there is good reason to doubt that they have anything more than an ordinal scale.

The question then is, are we coming to illusionary conclusions due to the use of inappropiate statistics? Theoretically the question is clear but pragmatically

it seems to be a question of opinion and personal discretion.

Upon glancing at the data of this study, several university statisticians were quick and confident in stating that this was interval data. Nevertheless, when one looks at the scales in a less mechanical fashion one has good reason to doubt that there is anything more than an ordinal scale. Given all of the above a cautious approach was taken. In many instances both parametric and non-parametric tests were used to allow for comparisons.

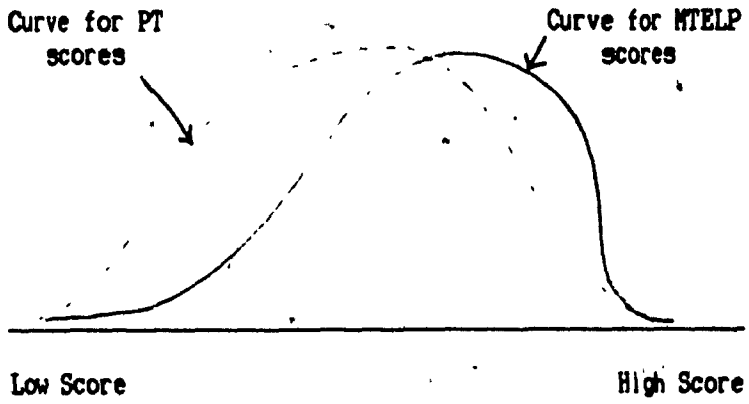
Distribution of the Test Sample for this Study

The data for this study is divided into two parts: the MTELP scores ranging potentially from 0 to 100, and the Pragmatic Task scores ranging from 1 to 5. Since the two tests were written by the same applicants it would seem to be reasonable to expect a similar distribution of scores for both tests. If the difference in distribution between the two tests were more than should be expected on the basis of chance variation and if such varience could not be attributed to the fact of having more than one sample, then the likely factor for the source of this variation would be the testing instruments. For example, there might be a

ceiling effect with the MTELP which would tend to cluster scores together which in the Pragmatic Task would be spread out presuming such a ceiling effect were not also occurring with the Pragmatic Task. If such were the case we might get distributions like those shown in Figure 3.

Figure 3

This Figure Shows a Possible Distribution for MTELP and Pragmatic Task Scores if There Were a Ceiling Effect with MTELP Scores.



Scores from the two tests might be rank-ordered similarly without their distributions being the same. For example, some applicants who have a score of two on the PT might all have MTELP scores which are lower than those obtained by the applicants who scored a three on the PT. In such a case the ordering of the ranks would be the same. Still, this being the case it is possible that at the same time that the distance between the MTELP scores for these applicants might be much smaller or larger than the distance between their PT scores. If such were the case we would not have equal intervals for the two tests and therefore the data would be

violating one of the premises for using parametric statistics.

Considering this problem a Chi-square test was performed. The test could not prove that the data had normal distribution but it could at least tell us wether or not the distribution of scores for the two tests were the same. One would certainly expect some variation between the distributions of scores for the two tests. The question would then be whether or not more variation would be found than could be expected on the basis of chance.

The Chi-Square Test for Goodness of Fit

First of all the composite scores for the two readers of each Pragmatic Task were calculated. Hence if one reader gave a score of 5 and the other reader gave a score of 4 the composite score was 4.5. The composite score was used instead of the score of one reader because the scores of two readers would give the best possible prediction of the true value of a Pragmatic Task. Of course even more readers would have been better, but the data only had two reader scores per Pragmatic Task. The composite scores yielded 9 possible observations, namely: 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. The frequency for each observation was

then calculated and the results obtained are shown in Table 22.

Table 22

Distribution of Readers' Composite Scores on the Pragmatic Task. There are Two Scores for Each PT and Four Readers Take Part in Making up All the Scores.

OBSERVATION LEVEL	FREQUENCY
5	6
4.5	8
4	9
3.5	17
3	14
2.5	14
2	15
1.5	3
1	4
$\bar{X}$ = 3.08	TOTAL 90

The MTELP scores were then rank-ordered and the mean for the MTELP scores was calculated. The MTELP scores were then divided into 9 observation levels so as to make them comparable to the PT scores. Since 3.08 was the mean for the PT scores the observation level of 3 (closest to 3.08) was used to fix the intervals for the MTELP.

For observation level 3 of the PT the frequency of scores was 14. So 14 was divided in two to get 7 and then looking at the rank-ordered MTELP scores 7 scores were counted up and down from the mean score for MTELP. This set the observations of scores from 71 on the MTELP to 81 on the MTELP as comparable to the observation of 3 on the PT. The other observation levels of the MTELP were then made, up and down from 71 and 81 and the frequency for each observation was counted. The results are seen in Table 23.

Table 23

Distribution of Readers' Composite Scores on PT and the Distribution of MTELP Scores on the Transformed Scale.

Observation levels for the PT	Frequency/ observation for the PT	Observation levels for the MTELP	Frequency/ observation MTELP
5	6	115 to 125	0
4.5	8	104 to 114	0
4	9	93 to 103	6
3.5	17	82 to 92	43
3	14	71 to 81	14
2.5	14	60 to 70	12
2	15	49 to 59	10
1.5	3	38 to 48	5
1	4	27 to 37	0
PT $\bar{X}$=3.08	total 90	MTELP $\bar{X}$=76.96	Total 90

Given the above frequencies per observation for the PT and the MTELP the mean frequency (between PT frequency & MTELP frequency) was taken for each observation to create the best possible prediction of the distribution of the sample given the two distributions from these two tests. Doing so yielded the numbers seen in Table 24.

Table 24

Frequency/Observation for the PT and the MTELP and the Best Possible Prediction of Expected Frquencies.

PT frequency/observation	6	8	9	17	14	14	15	3	4
MTELP	0	0	6	43	14	12	10	5	0
Expected	3	4	7.5	30	14	13	12.5	4	2

Using the above frequencies the Chi-square test with 16 degrees of freedom was performed. The middle observations for the PT and MTELP cannot be counted for degrees of freedom because in transforming the MTELP scale these two observations were fixed. The results of the Chi-square test are as seen in Table 25.

Table 25

Chi-square Goodness of Fit Test Results Comparing PT and MTELP Frequencies with Expected Frequencies.

	f_0	f_e	diff.	$\frac{(f_0 - f_e)^2}{f_e}$
	6	3	3	3.0
	8	4	4	4.0
	9	7.5	1.5	0.3
PT score	17	30	13	5.6
	14	14	0	0.0
Distribution	14	13	1	0.1
	15	12.5	2.5	0.5
	3	4	1	0.3
	4	2	2	2.0
	0	3	3	3.0
	0	4	4	4.0
	6	7.5	1.5	0.3
MTELP score	43	30	13	5.6
	14	14	0	0.0
Distribution	12	13	1	0.1
	10	12.5	2.5	0.5
	5	4	1	0.3
	0	2	2	2.0
				$\chi^2 = 31.6$

Conclusion

"The hypothesis we are testing is that the observed frequencies do not differ from the frequencies

that are expected in each category (null hypothesis)." (Kerlinger, 1964, p. 210) Looking at the table for 16 degrees of freedom we find that for a p(probability) of .02 we have an x of 29.63. Our x is greater than 29.63. If the null hypothesis were true then the probability of obtaining our distribution of scores would be less than 1 in 50. Therefore we must reject the null hypothesis and conclude that the observed frequencies differ significantly from the expected frequencies.

Appendix F

How Representative is the Sample of this Study?

The sample for this study (n=90) was obtained from a group of applicants wishing to become students at Concordia University. These applicants wrote the CELDT battery in March 1980. The question arises of how representative the sample is of university students in general and of Concordia University students in particular. For the former we cannot say but as to the latter the registration figures at Concordia University for the fall of 1980 were used to compare to the results for the sample. Due to the short questionnaire the applicants filled out before writing the CELDT battery it was possible to compare the applicants of this study with the student body as a whole on the basis of sex and degree sought.

Concordia University has many part-time students, in fact slightly more than 50% of the students registered in the fall of 1980 were part-time. In comparing the sample of this study with university registration it was decided to include the figures for part-time students because it is possible that amoung the applicants for this study that there were both people who wanted to study full-time and part-time.

Table 26

Comparison of Male/Female Distribution for the Sample from this Study and for the 1980 Registration at Concordia University

Sex	# for study	% for study	# for university registration	% for university registration	% diff.
female	48	53.3	-	48.6	4.7
male	42	46.7	-	51.4	4.7

* For university registration, only percentage was available

Table 27

Comparison for Degree Sought as Indicated for the Sample from this Study and for the 1980 Registration at Concordia University

Degree sought	# for study	% for study	# for univ. registration	% for univ. registration	% difference
Arts	23	25.6	7480	36.7	11.1
Education	8	8.9	135	0.7	8.2
Science	4	4.4	1801	8.8	4.4
Commerce	11	12.2	5768	28.3	16.1
Administration	5	5.6	754	3.7	1.9
Engineering	12	13.3	1285	6.3	7.0
Computer Sc.	9	10.0	811	4.0	6.0
Fine Arts	18	20.0	2329	11.4	8.6

Using the above results a Chi-Square was done so as to compare the distributions of the Concordia University student body registration for the fall of 1980 and the distribution for the applicants who wrote the CELDT battery from whose scores the data for this study was obtained. The distribution in both cases was based on degree sought. The results are as seen in Table 28.

Table 28

Chi-square Test with Registration of 1980 as the Expected Distribution and the Distribution for this Study as the Obtained Distribution with Degree Sought as the Basis for the Distribution. The Scores for Both Distributions are in Percentages.

Degree sought	f_o	f_e	difference	$\frac{(f_o - f_e)^2}{f_e}$
Arts	25.6	36.7	-11.1	3.3
Education	8.9	0.7	8.2	96.0
Science	4.4	8.8	-4.4	2.2
Commerce	12.2	28.3	-16.1	9.1
Administration	5.6	3.7	1.9	1.0
Engineering	13.3	6.3	7.0	7.8
Computer Sc.	10.0	4.0	6.0	9.0
Fine Arts	20.0	11.4	8.6	6.5
DF = 8				$X^2 = 134.9$

For 8 degrees of freedom the table shows that for a x greater than 26.12 p=.001. Hence we must reject

the null hypothesis and presume that the distributions are significantly different.

It is important to note the difference in distribution (in terms of degree sought) for the sample from this study and the Concordia University student body in general. Still, it is worth noting that there are some reasons for not expecting the distributions to be similar. The registration figures refer to all Concordia students and not only to non-anglophone students. The sample from this study deals only with non-anglophones and it might be questioned as to whether or not it is valid to expect the non-anglophone students to have the same academic interests as the student body at large.

C. Native Language:

Table 20

Native Languages of the Applicants Used for this Study

Native Language	Number of Applicants
French	38
Arabic	9
Spanish	7
Vietnamese	7
Iranian	4
Italian	4
Greek	3
Armenian	2
Chinese	2
German	2
Indonesian	2
Bengali	1
Czech	1
Dutch	1
Hebrew	1
Polish	1
Russian	1
Tagalog	1
Telugu	1
Turkish	1
Urdu	1

D. Choice of Degree: Before writing the CELDT battery the applicants are asked to fill a short questionaire. One of the sections asks the question 'Degree sought'. The applicants have eight possible choices. Their choices are shown in Table 21.

Table 21

The Number of Applicant Responses for Each Catagory of the Question 'Degree Sought'

Catagory	Number of Responses
Administration	5
Arts	23
Commerce	11
Computer Science	9
Education	8
Engineering	12
Fine Arts	18
Science	4

E. Sex: The ninety subjects are composed of forty-eight females and forty-two males.

F. Age: The average age of the applicants is 24.2 years old.