PREDICTIVE VALIDITY OF TOEFL SCORES ON FIRST TERM'S GPA AS THE CRITERION FOR INTERNATIONAL EXCHANGE STUDENTS

by

ZHENG YAN

M.Ed., Northeast Normal University, 1986

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

Department of Language Education

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

March, 1995

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without written permission.

Department of Language Education

The University of British Columbia

Vancouver, Canada

Date June 12, 1995

Abstract

The Test of English as a Foreign Language (TOEFL) has been used in making admission decisions for over 30 years; however, the predictive validity of the test has been uncertain. The present study was intended to investigate the predictive validity of TOEFL scores on first term's grade point average (GPA). Participants were 97 second-year university students, 46 male and 52 female, in an international academic exchange program. Most majored in Humanities and Social Sciences. The predictor variables in the study included TOEFL total scores, TOEFL section I scores, TOEFL section II scores, TOEFL section III scores, oral proficiency interview scores, writing sample scores, and gender. First term's GPA was the criterion variable. The data were analyzed by multiple regression analysis with a hierarchical procedure. The results were interpreted on the basis of Cohen's (1988) conventional definitions on the effect size of R^2.

The main findings of the study indicate that: (a) TOEFL total scores have a medium level of predictive validity on GPA (ΔR^2=.142, p<.001); (b) TOEFL section I scores have a medium level of predictive validity (ΔR^2=.044, p<.05); (c) TOEFL section II scores have a medium level of predictive validity (ΔR^2=.112, p<.001); (d) TOEFL section III scores have a negligible level of predictive validity (ΔR^2 =.005,

$p>.05$); (e) Oral proficiency interviews scores have a negligible level of predictive validity ($\Delta R^2 = .010$, $p>.05$); (f) Writing samples scores have a small level of predictive validity ($\Delta R^2 = .047$, $p<.05$); And (g) gender has a medium level of predictive validity ($\Delta R^2 = .130$, $p<.001$). The findings of the study thus validate the use of TOEFL scores as one of the requirements for admission in the international exchange program and provide new empirical evidence for investigation of the relationship between language proficiency and academic achievement.

TABLE OF CONTENTS

List of Tables

List of figures

Acknowledgments

The growth of a flower, no matter how small, has to appreciate the Sun's enlightenment, the Rain's refreshment, and the Mother Earth's nutrition and grounding. Here, I wish to extend my heartfelt thanks to:

Dr. Lee Gunderson, my MA program advisor, for giving over three years of tireless guidance;

Dr. Richard Berwick and Dr. Stephen Carey, my research committee members, for their invaluable support and advice;

Dr. Areigh Reichl for providing consultation on statistics and William McMichael for providing consultation on the UBC/Ritsumeikan program. They both read the whole thesis and gave constructive criticism;

Sheri Wenman, Jean Hamilton, and the UBC/Ritsumeikan program's instructors and students for their constant support and cooperation for my thesis project;

Dr. Robert Kantor, Director of ETS's TOEFL Program, for providing both professional consultation on TOEFL and a long list of ETS' free publications;

Dr. Xiufeng Liu, Dr. Dean Mellow, Bingzheng Liu, and particularly Dr. Nand Kishor for their inspiration in the development of my thesis project;

Victoria Dixon, Lynda Hayward, and Elizabeth Crittenden for their laborious proof-reading on the different chapters of the thesis;

Dr. William Mackey, Dr. David Robitaille, Dr. Karen Armstrong, Dr. Robert Conry, Dr. Bernard Mohan, Dr. Marshall Arlin, Dr. Jon Shapiro, Dr. Marion Crowhurst, and Dr. Judith Johnston for their inestimable support, teaching, and/or encouragement;

My friends, White Harvey, Cathy Galloaher, Cuhui Zhao, Elizabeth Smith, Gary and Mary Gates, Jingzi Wang and Dr. Yuan Gao, Prof. Kunwei Wang, Dr. Lianqin Wang, Jim and Katherine Yuen, Zhong Liu, Dan Zhang and Yaoyao, Dr. Glen Dixon and Victoria Dixon, Roberta Buck, and Dr. Leigh Faulkner for their immeasurable support;

My parents and brother for their deep understanding of and exhaustive financial support for my study at UBC;

And my wife, Jingkai Zhang, for sharing my stress and happiness.

Chapter One

Introduction

This chapter presents the research problem under study. The specific research questions and detailed definitions of terms are also given.

Research Problem

The Test of English as a Foreign Language (TOEFL) is the most widely used test of English as a foreign language in the world. It was first administered in 34 countries in 1964 (Oller & Spolsky, 1979). At present, as reported by the Educational Testing Service (ETS), TOEFL is given on a monthly basis at over 1,200 test centers in 175 countries or regions around the world, with a population of approximately 700,000 examinees every year (ETS, 1994a, 1994b).

The primary function of TOEFL, as stated in the latest TOEFL Test and Score Manual, is "to measure the English proficiency of international students wishing to study at colleges and universities in the United States and Canada" (ETS, 1992, p. 6). Although considerable evolution of TOEFL has occurred during its 30 years of development, the primary function has never changed. TOEFL scores are currently required for admission into undergraduate or graduate programs by more than 2,500 colleges and universities in the USA and Canada (ETS, 1994c).

A great deal of research has been conducted to validate the use of TOEFL (Hale, Stansfield, & Duran, 1984; ETS, 1994d). A large proportion of the research has explored

TOEFL's predictive validity with grade point average (GPA) as the criterion. Since English proficiency is necessary to achieve academic success in an English environment, there should be a positive relationship between English proficiency and academic achievement, and consequently a positive relationship between TOEFL scores as an indicator of English proficiency and GPA as an indicator of academic achievement. Accordingly, TOEFL scores should have strong predictive validity in predicting GPA; however, TOEFL prediction studies have consistently revealed widely divergent results (Graham, 1987; Hale, Stansfield, & Duran, 1984). Although researchers generally agree that English language proficiency is important for academic achievement, they have not yet been able to reach a consensus on TOEFL's predictive validity.

The problem is, therefore, that TOEFL has been used worldwide by thousands of institutions to make admission decisions for 30 years, but the predictive validity of TOEFL is still an unsolved question for professionals in language education.

Research Questions

This study was designed to estimate the predictive validity of TOEFL scores on GPA for students in the 1993-94 UBC/Ritsumeikan Academic Exchange Program, which was jointly administered by the University of British Columbia (UBC) of Canada and Ritsumeikan University of Japan. As an institutional validity study, it was intended to provide

empirical evidence to investigate whether TOEFL scores predict GPA, and to explore how language proficiency is related to academic achievement.

The study addressed the following specific research questions:

1. Do TOEFL scores predict GPA for international exchange students?
2. Do grades measuring English writing and speaking abilities predict GPA for international exchange students?
3. Do non-language variables predict GPA for international exchange students?

Definitions of Terms

Predictive validity. Validity refers to the appropriateness of inferences from test scores or other forms of assessment (American Psychological Association, 1974, pp. 25-27). Based upon the kinds of inferences one might wish to draw from test scores, people traditionally refer to the following types of validity: criterion-related validity, including both predictive validity and concurrent validity, content validity, and construct validity. Predictive validity indicates the extent to which one can predict future performances from prior information.

Predictive variables. The information that is used to make a prediction is typically referred to as a predictive variable or simply as a predictor.

Criterion variables. The event or outcome to be predicted is typically referred to as a criterion variable or simply as a criterion.

GPA. This is an acronym for grade point average. It is used as a measure of academic achievement in subjects or courses, usually obtained by dividing the sum of the total grade points by the total number of courses. In the current study it is used as an indicator of university academic achievement.

TOEFL. This is an acronym for the Test of English as Foreign Language. The current study uses TOEFL scores as indicators of English Language proficiency.

Model. A model is a hypothesized structure used for the investigation of interrelations between variables or hypotheses. After variables have been identified, or hypotheses have been advanced in the course of inquiry, it may be necessary to advance a model that provides a structure for the interrelations between the set of variables or hypotheses. Model building and model testing are two strategies that can be employed in inquiry. Both correlation and regression can contribute to model building (See Husen, 1994, pp. 3865-3873).

Language proficiency. This term means progress towards the attainment of a high degree of knowledge and skill in English language. In the present study, this is used intentionally to distinguish it from language competence, language performance, and language aptitude.

Academic achievement. In this study, academic achievement refers to performance by students in academically oriented courses. It is interchangeable with academic success.

Chapter Two

Literature Review

Introduction

The review of literature in this chapter focuses on research findings related to the predictive validity of TOEFL scores with GPA as a criterion variable. The review is divided into two parts: factors that influence academic achievement, and factors that affect the estimation of TOEFL's predictive validity. Part I examines conceptual issues in studies of the predictive validity of TOEFL scores, while Part II concentrates on methodological issues.

Background

The Test of English as a Foreign Language is a standardized test which uses a multiple-choice format to evaluate the English language proficiency of non-native speakers. Between 1963 and 1976, TOEFL contained five sections: Listening Comprehension, English Structure, Vocabulary, Reading Comprehension, and Writing Ability. Since September of 1976, TOEFL has consisted of three sections: Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension. The two forms of the test differ in testing items used and testing time allowed, but the score scale of both tests is the same. The different sections of the test were designed to measure different language skills within the general domain of language proficiency. Three decades of testing administration and extensive research have shown that TOEFL

has a high degree of reliability and validity (ETS, 1992, pp. 30-36).

Part I: Factors Influencing Academic Achievement

The first part of the literature review presents both a two-level conceptual structure in the study of TOEFL's predictive validity and a five-level classification scheme of factors influencing academic achievement. It will then examine both language factors and non-language factors related to academic achievement.

A conceptual structure

TOEFL scores are but one indicator of language proficiency, while GPA is but one indicator of academic achievement. In a sense, the relationship between TOEFL scores and GPA is a surface-level manifestation of the parallel but underlying relationship existing between

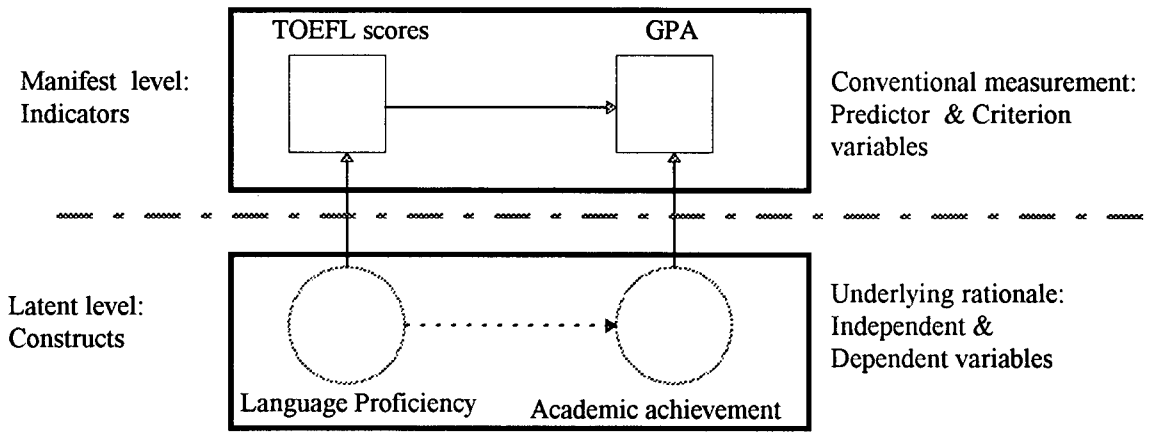

Figure 2.1. A two-level conceptual structure in the study of TOEFL's predictive validity.

language proficiency and academic achievement. This relationship is illustrated in Figure 2.1.

It is shown that structurally a TOEFL prediction study has two portions: the manifest level and the latent level. This is analogous in structure to an iceberg, its visible part being above sea level and the rest below sea level. The manifest-level portion is a conventional statistical measurement of the relationship between TOEFL scores as a predictor and GPA as a criterion. The latent-level portion is a theoretical assumption about the relationship between language proficiency as an independent variable and academic achievement as a dependent variable.

Figure 2.1 also demonstrates that these two portions are not separate from but harmonize with each other. In a TOEFL prediction study, the underlying theoretical assumption about the relationship between language proficiency and academic achievement should justify statistical methods used to measure TOEFL's predictive validity, while the conventional measurement of the relationship between TOEFL scores and GPA should fit the underlying theoretical rationale. For every study, in fact, the research method used ought to match well with the proposed theoretical assumption. For example, one might conduct a correlation study to analyze the relationship between children's IQ and the size of shoes they wear. However, this study would not make any sense because the

statistical method in the study, no matter how well it would be utilized, lacks a logical supporting rationale.

Clearly, a good TOEFL prediction study needs a proper statistical approach, but more important, it largely relies on a strong rationale. This is simply because the hypothesized relationship between language proficiency and academic achievement influences how the relationship between TOEFL scores and GPA is measured. Unfortunately, the issue of the underlying rationale for the TOEFL prediction study has been repeatedly ignored. Thus, in proposing such a two-level conceptual structure for TOEFL prediction studies, the intention is to emphasize the importance of a comprehensive examination of factors influencing academic achievement. This examination serves to establish a solid rationale underlying the measurement of the predictive validity of TOEFL scores.

A classification scheme

Numerous studies have documented a great variety of factors influencing academic achievement, such as intelligence, language, motivation, personality, interest, age, teacher expectation, ethnicity, learning style, teaching strategies, family involvement, classroom environment, and peer pressure. Since it is nearly impossible to list all of these factors within a limited space, the present study grouped them into a five-level classification scheme sequenced from external factors to internal factors. This classification is preliminary and

heuristic. It is used to show the hierarchical structure and complicated interrelations of the various factors related to academic achievement (See Figure 2.2).

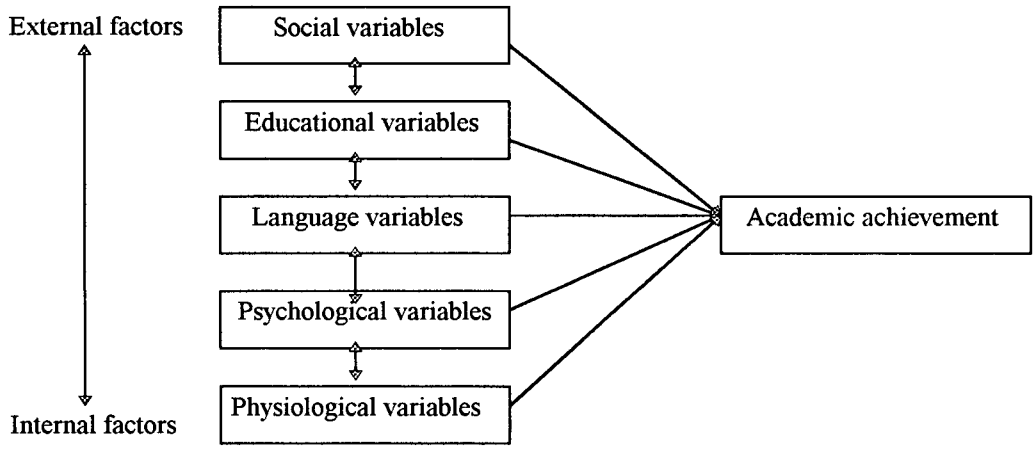

Figure 2.2. A five-level classification scheme of factors influencing academic achievement.

The first major type of variable influencing academic achievement is social. This category includes social development, socioeconomic status, cultural background, ethnicity, social attitude, family environment, parental involvement, morals and values, marital status, employment chances, and religious beliefs.

The second major type of variable influencing academic achievement is educational. Examples include curriculum implementation, educational objectives, instructional materials, teaching approaches, characteristics of students, characteristics of teachers, classroom interactions, time

spent on learning, prior knowledge, learning style, teacher expectation, school assessment and evaluation, subject matter, students' status, and classroom environment.

The third is a linguistic category with such variables as first language(L1), second language(L2), bilingualism, reading, speaking, listening, writing, genre, language proficiency, communicative competence, receptive skills, productive skills, vocabulary, and meta-awareness of language.

The fourth category consists of psychological variables such as motivation, cognition, emotion, personality, attention, attitude, interest, aptitude, anxiety, creation, temperament, and self-esteem.

Physiological variables represent the fifth category, including gender, genetic factors, maturation, fitness, brain lateralization, aging, health, and nutrition.

This classification scheme reveals how a large numbers of factors may contribute to academic achievement. Language factors are only one group among five which influence academic achievement. Non-language factors, such as social variables, educational variables, and physiological variables, also play important roles. Oversimplification of the process of academic learning or the interrelationship among factors concerned may lead to erroneous findings. Thus, the following literature review is organized into two sections: first, language factors and academic achievement, with a focus on the relationship between second language

proficiency and university academic achievement; and second, non-language factors and academic achievement, with a highlight on the relationship between gender differences and academic achievement.

Language factors and academic achievement [1]

It has been generally recognized that language is the major medium of learning (Mohan, 1986) and language proficiency is important to academic success. For those who study in educational institutions where the language of instruction is their second language, in particular, their L2 proficiency remarkably affects, even determines, academic achievement. However, research in L2 education shows that the strength of the relationship between L2 proficiency and academic achievement varies for different language skills and across content areas.

L2 skills and academic achievement. Cummins (1981) described two types of language proficiency: Basic Interpersonal Communication Skill (BICS) and Cognitive Academic Language Proficiency (CALP). He pointed out that academic language proficiency, rather than daily conversational competence, is necessary for academic success. His findings have been supported by many empirical studies (Collier, 1987). Other researchers have explored the

[1] This review mainly focuses on studies concerning the relationship between second language proficiency and academic achievement. This is because of the limitation of the space, and the topic being too broad to cover. More important, it relates directly to the present research questions.

relationship among different language skills, listening, speaking, reading, and writing, to academic achievement.

Johns (1981) conducted a study involving an academic skills questionnaire with 200 faculty from all departments at an American university in order to determine which language skills among reading, writing, speaking, and listening were most essential to non-native speakers' success in their university classes. Results of the study showed that receptive skills, both reading and listening, were ranked first.

Ostler (1980) reported similar findings in a study of a group of ESL college students' assessment of what academic skills they needed to achieve academic success. The study revealed that academic reading skill was ranked as the most needed among sixteen language skills. Other highly ranked skills were taking notes, asking questions in class, reading journals, and writing research proposals.

In a study of 178 university professors' judgments of errors in the writing of non-native speaking students, Santos (1988) reported that professors seemed to place more emphasis on language features than on content features, and lexical errors in writing were rated as the most serious. This suggested that basic knowledge of vocabulary in writing plays an important role in academic achievement.

Magan (1986) conducted research on the relationship between speaking proficiency and academic achievement of 40 college French students. His findings revealed a significant

positive relationship between speaking ability and academic success.

In a canonical correlation analysis, Ho and Spinks (1985) found that listening ability was not as predictive of academic performance at the university level as were speaking, reading, and writing abilities. They argued that it was likely that listening difficulties might be compensated through additional reading.

The foregoing research findings suggest that different language skills have different impacts on academic achievement, although it appears that no consensus exists yet in terms of which language skill plays the most important role.

L2 proficiency across subject matters. Mohan (1986) analyzed the relationship between language and content and considered the nature of language in education as a medium of learning. Mohan's theoretical perspective provided insight into the relationship between second language proficiency and academic achievement in different subject areas across the curriculum.

Slark and Bateman (1981) studied non-native English speakers' college academic achievement. Their findings showed that there was a significant positive correlation between language scores and course grades in two courses (Anthropology and Sociology), whereas three other courses (Chemistry, Mathematics, and Music) consistently showed negative correlation coefficients. The results indicated

that courses in social sciences required higher levels of language proficiency than those in natural sciences and music.

Crandall and others (1987) analyzed the relationship of ESL language development to academic achievement in mathematics, science, and social studies. They argued that although the exact relationship between ESL language development and content learning of these subjects was not clearly understood, both a minimal level of language proficiency with specific linguistic registers and a minimal knowledge of the academic area were required for academic success.

As far as mathematics learning is concerned, studies with monolingual English speakers have revealed a high positive correlation between mathematics achievement and English reading ability (Aiken, 1971; Duran, 1979). These results are interesting because mathematics uses its own symbolic system except for word problem solving. In MacNamara's studies (1966, 1967), bilingual children kept pace with monolinguals in mechanical arithmetic, but fell behind in solving word problems. Several researchers have found that language minority students frequently do not understand the language used to present mathematics test problems (DeAvila & Havassy, 1974; Moreno, 1970).

In short, research findings show that there is a relationship between language factors and academic achievement for different language skills and in different

subject areas, but do not reveal identifiable patterns. As Vinke and Jochems (1993) pointed out, there is no generally acknowledged theory on the precise nature of the relationship between language proficiency and academic achievement. Therefore, making conclusive statements about the relationship is premature.

Non-language factors and academic achievement

Comprehensive studies on non-language variables affecting academic achievement. Many researchers have examined the effects of non-language factors, individually or in combination, on academic achievement. These factors include teacher expectation (Rosenthal and Jacobson, 1968), achievement motivation (Ames & Ames, 1984), home environment (Soto, 1990), and social disadvantage (Ushasree, 1990). In addition, comprehensive studies on varied factors affecting academic achievement have been conducted in order to identify factors that significantly and consistently influence academic achievement and to provide empirical evidence about weights and interrelationships among these factors.

Ho and Spinks (1985) examined the effects of four variables, verbal intelligence, English language skills, personality, and attitude, on university academic performance. Their findings showed that (a) English language skills had the most predictive value, accounting for about 10% of the variance in academic performance; (b) Verbal

intelligence, attitude (excepting study orientation) and personality were not predictive of academic performance.

Walberg, Schiller, and Haertel (1979, 1982) collected and analyzed the review literature of the 1970s on the effects of instruction and related factors on cognitive, affective and behavioral domains. Based on a synthesis of 23 major research topics addressed by thousands of studies, they found that nine variables appeared to have consistent causal influences on academic leaning: student age or developmental level, ability, motivation, amount of instruction, quality of instruction, the psychological environments of the class, home, peer group outside school, and exposure to the mass media.

By performing a linear structure relation analysis (LISREL), Walberg and three other co-researchers (1984) compared five causal models to examine the relationship between achievement in science and a combination of eight variables. The eight variables were students' ability, home environment, peer group, exposure to mass media, social environment, time on task, motivation, and instructional strategies. Results showed that among the eight factors students' ability (r ranged from .72 to .75) and motivation (r ranged from .11 to .12) consistently had the largest influences on science achievement.

In another research synthesis (Walberg, Pascarella, Haertel, Junker, & Boularger, 1982), 14 major variables which affect academic achievement in science, math, social

studies, and reading were listed. The 14 variables were age, achievement, attitude, socioeconomic status, quality of instruction, quantity of instruction, education, home, peer, homework, media-TV, extracurricular, stimulation, and gender.

Gender differences and academic achievement. Numerous studies have discussed gender differences and academic achievement. Maccoby and Jacklin (1974) in their widely cited book summarized and analyzed a large amount of research on gender differences and concluded that: (a) Girls have greater verbal ability than boys; (b) Boys excel in visual-spatial ability; (c) Boys excel in mathematical ability; And (d) males are more aggressive. Their results were supported by findings of large scale studies conducted nationally or internationally.

The National Assessment of Educational Programs (Husen, 1994, pp. 5425-5426) in its large scale studies over ten years found that the girls performed consistently better on both reading and writing tests than boys, but not on science.

The International Association for the Evaluation of Educational Achievement (IEA) studies of mathematics and science (Keeves, 1973) showed that, while the general pattern of results was one of superior performance by male students in both subjects, there was considerable variation between countries in the extent to which boys exceeded girls in performance.

Walker (1976, in Husen, 1992, p. 5426) reported another IEA study on gender differences in six subjects areas: reading, literature, English as a foreign language, French as a foreign language, and civic education. On reading comprehension tests, boys showed lower performance than girls in a majority of countries, but in general these differences were slight. On the literature tests, in all countries the boys did less well, and they also showed less interest in literature. Again, in a study of the teaching of English as a foreign language, the boys scored below the girls on both the reading and listening tests, but the differences were small. In a study of the teaching of French as a foreign language, statistically significant gender differences in the learning of French were recorded in English-speaking countries, with girls performing better than boys. In civic education achievement tests, the boys generally recorded higher scores than girls.

Several studies examined issues of gender differences in language tests. Landsheere (1994) found that boys perform marginally better than girls on multiple-choice tests and problem-solving exercises. Girls perform better than boys on essay tests in written composition and are generally assigned higher grades in school-based assessments. In another study (Zeidner, 1987), the researcher analyzed the English language aptitude test scores of 824 full time Jewish students in Israel and found that a small degree of gender differences in test scores was observed, tending to

overpredict the first year's GPA of males and underpredict that of females. The researcher argued that this might be the result of differential grading practices and unevenness in the number of males and females in courses, rather than as a fact of nature.

In summary, much research has documented gender differences in academic achievement in such subject areas as mathematics, science, social studies, language arts, and foreign languages. It appears clear that (a) there are gender differences in academic achievement; (b) these differences should not be exaggerated; and (c) many factors contribute to gender differences. In fact, gender should not be considered as a purely biological entity, but rather, a composite variable combining physiological, psychological, and sociological components. Gender differences in academic achievement originate from a variety of sources, such as participation differences, abilities differences, biological differences, socialization differences, differences in attitudes and their effects, and differences in the expectancy of success (Husen, 1982, pp. 5428-5430).

Part II: Factors Influencing TOEFL's Predictive Validity

The following part of literature review examines five major methodological factors which substantially influence the estimation of TOEFL scores' predictive validity. These factors are: (a) Which analytical models are employed? (b) What subject variables are involved? (c) What predictor

variables are used? (d) What criterion variables are selected? And (e) how results are computed and interpreted?

Analytical models [2]

An analytical model refers to a hypothesized structure to emulate and analyze the interrelations between variables. There are many analytical models used for prediction or explanation studies (Pedhazur, 1982). It is important to choose and employ appropriate analytical models in conducting a study of TOEFL's predictive validity. The model should be chosen properly in order to fit the data as well as the research question under study. It should be used correctly in order to meet the assumptions underlying the model.

The correlation model versus the regression model. Most studies estimating TOEFL's predictive validity have applied the correlation model as the sole analytical model (Chase & Stallins, 1966; Abdzi, 1967; Kwang & Dizney, 1970; Martin, 1971; AACRAO, 1971; Pack, 1972; Heil & Aleamoni, 1974; Shay, 1975; Harcey, 1979; Bostic, 1981; Riggs, 1982; Odunze, 1982; Light, Xu & Mossop, 1987; Johnson, 1988; Light & Wan, 1991; Ayers & Ouattlebaum, 1992). These studies usually estimated TOEFL's predictive validity by calculating correlation

[2] The present study purposely used the term of analytical models instead of statistical methods or statistical models. A scientific analysis is not identical to a statistical method. Even for quantitative research in which the statistical method is its essential component, the statistical method cannot cover all the content that the analytical model contains, such as model construction and model modification.

coefficients between TOEFL scores and GPA. The correlation model has dominated TOEFL prediction research.

Some researchers (Schreder & Pitcher, 1970; Sharon, 1972; Gue & Holdaway, 1973; Stove, 1982; Hassan, 1982; Yule & Hoffman, 1990) have used the correlation model as the main analytical model with the regression model as a supplement. These authors estimated correlation coefficients ($\underline{r}$) and proportion of variance accounted for by regression ($\underline{R}^2$), in some cases with regression coefficients ($\underline{b}$ & β) or the regression equation).

A few studies have adopted the regression model as the main analytical tool with the correlation model as its integrated component (Wilcox, 1975; Andalib, 1976; Ayers & Peters, 1977; Sokari, 1981). In these studies, correlation coefficients are calculated as one of the basic descriptive estimates. The main procedure is to perform a regression analysis so that the regression equation, squared multiple correlation, and/or regression coefficients are obtained.

Which analytical model should be chosen for prediction studies? This issue has been discussed extensively in psychometric research since the 1950s (Kendall, 1951; Fish, 1958; Binder, 1959; Ezekiel & Fox, 1959; Fox, 1968; Warren, 1971; Thorndike, 1978;). Based on these studies, Pedhazur (1991) concluded that when the focus of the research is on the explanation, or the prediction, of dependent variables, the regression model is appropriate (p. 409). In TOEFL prediction studies, the research purpose is to see how well

TOEFL scores predict the criterion variable GPA, but not to describe the association between two arbitrarily selected variables. Thus, the regression model rather than the correlation model is the proper solution.

The simple regression model versus the multiple regression model. TOEFL prediction studies using the regression model as their main or supplemental analytical tool can be classified into three groups in terms of the number and the variety of predictor variables involved.

In the first group of studies, only one predictor variable is used in the regression model (for instance, Hassan, 1982).

In the second group, multiple predictor variables of English language proficiency are used as predictors in the regression models (for example, GRE-V and MTELP scores, Abdzi, 1967; TOEFL's five-section scores, Sharon, 1972; Prior- and post-admission TOEFL scores, and interview scores, Gue and Holdaway, 1973; TOEFL's overall and sectional scores, Hu, 1991).

Multiple predictor variables with multiple features are used in the third group of studies (for example, TOEFL and LSAT. Schrader & Pitcher, 1970; TOEFL, ACT, SAT, high school GPA, and age. Andalib, 1976; TOEFL, ESL course grades, native language, major areas of study. Stove, 1982; TOEFL, GRE-V and GRE-Q. Yule & Hoffman, 1990).

The number and the variety of predictors in the regression model are dependent upon the complexity of the

research problem under study. When a one-cause-one-effect relationship exists, a simple regression model should be used to predict a phenomenon completely determined by a single factor. For more complicated phenomena, more predictors are needed. For phenomena influenced by different types of factors, a multiple regression model with different types of predictors is required. In social and educational research, the multiple regression model is necessary in most cases to make the prediction study defensible.

There are manifold factors affecting academic achievement, therefore, a multiple regression model with multiple predictors is appropriate to predict GPA. Using only one predictor, or the language-based predictors, makes it difficult to gain an accurate prediction of GPA. Many TOEFL prediction studies, as reviewed above, used language-based variables; as a result, they frequently obtained relatively smaller $\underline{R}^2$, even though more similar language-based predictors were added into the regression equation. Thus in TOEFL prediction studies, we should not only choose multiple predictors, but also take into account the degree of diversity of the predictors.

More complicated models, such as path analysis model, Linear Structural Relations model, Hierarchical Linear Model, canonical analysis model, and discriminant analysis model, can also be used to analyze the complex relationship of factors affecting academic achievement.

The single-step regression calculation versus the comprehensive regression analysis package. Regression analysis should not be seen as the sole calculation of R^2, or of regression coefficients. An analytical process of the multiple regression typically involves integrated components and relevant techniques, including the checking of assumptions, detecting outliers (by using residual analysis and influence analysis), regression estimation, hypothesis testing, as well as power analysis (Cohen, 1988; Husen, 1994, p. 3866; Pedhazur, 1991). The procedures and techniques mentioned above examine the fit of regression models to data, the existence of outliers, the weighting of the variables, and the degree to which results can be generalized so that the quality of a multiple regression analysis can be optimized.

Multiple regression studies of TOEFL's predictive validity usually report R^2, the regression equation (Schreder & Pitcher, 1970; Sharon, 1972; Sokari, 1981; Stove, 1982; Yule & Hoffman, 1990; Hu, 1991), results of the stepwise regression (Gue & Holdaway, 1973; Andalib, 1976; Ayers & Peters, 1977), and/or standard error of estimation (S_{est}) and shrinkage (Hassan, 1982). However, it appears that few researchers, if any, perform the comprehensive regression analysis mentioned above.

Subject variables [3]

Many studies have reported that various subject variables affect the estimation of TOEFL's predictive validity (Hale, Stanfield, & Duran, 1984). These subject variables can be grouped into four categories.

(a) Personal information, such as gender, age, parents' educational level.

(b) Social factors, including native language, home country or region, citizenship, ethnic group, social adjustment, and occupation in home country.

(c) Academic background, for instance, areas of study, type of degree sought, educational level, and previous grades.

(d) Test-related information, such as TOEFL repeaters or non-repeaters, TOEFL scores in the Friday program or the Saturday program, and the like.

The following discussion, however, focuses on two issues related to subject variables. These issues cause serious problems but were often ignored in the estimation of TOEFL's predictive validity.

Sample size. Sample size is associated with the homogeneousness of subjects under study. Differences in

[3] In some studies, some subject variables were used as predictor variables, functioning as a moderator or mediator along with TOEFL scores to predict GPA. This could be also viewed as an evidence of influences of subject variables on TOEFL/GPA relation. It is this kind of unintended and easily-neglected effect of subject variables that make non-experimental research, including the TOEFL/GPA study, more complicated.

sample size have different effects on the predictive validity of TOEFL scores. Although almost all studies on the predictive validity of TOEFL scores reported their sample sizes, the range in sample size among varied from 15 to 900. Some TOEFL predictive studies lacked sufficient sample size (Bostic, 1981; Hassan, 1982; Riggs, 1982). Most studies used the cumulative sample size obtained across years (e.g., Schreder & Pitcher, 1970. n=63, from 1964 to 1969; Sharon, 1972. n=973, 1964-69; Pack, 1972. n=402, 1960-72; Gue & Holdaway, 1973. n=123, 1967-70). This kind of cumulative sample size might result in problems regarding the predictive validity of TOEFL scores. It might confound various subject variables, ignore the differences in the two forms of TOEFL (i.e., three-section and five-section), or lose unique information in sub-samples for each year.

Mean TOEFL scores. Mean TOEFL scores indicate the average level of English language proficiency of the subjects under study. They substantially influence the extent to which TOEFL scores predict academic achievement.

Wilcox (1975) found that one group of subjects with better ESL proficiency showed no relationship between TOEFL scores and GPA, whereas another group with lower English levels showed a significant relationship. He explained that English ability and academic success may be related at low levels of proficiency but unrelated at levels above certain threshold values. Wilcox's findings suggest that the existence of certain thresholds of TOEFL scores probably

results in a nonlinear relationship between English proficiency level and academic achievement.

Similarly, Johnson (1988) found that when English proficiency is relatively low, TOEFL scores can predict academic performance. With higher language proficiency, other variables such as prior exposure to subject matter, motivation, study skills, cultural adaptability, and even financial security, may became more important.

The TOEFL Test Manual (ETS, 1992) states that if the standard for English language proficiency is set at such a high level that only applicants with good English skills are admitted, there may be little relationship between TOEFL scores and any of the criterion measures. Because there will be no large variance in English proficiency among the group members, variations in success on the criterion variables will be due to other non-English causes. On the other hand, if the standard is set at too low a level, a large number of applicants selected with TOEFL scores may be unsuccessful in the academic program. There will be a relatively high correlation between their TOEFL scores and its criterion measures. Thus, with a standard that is neither too high nor too low, the correlation between TOEFL scores and subsequent success will be only moderate.

Mean TOEFL scores also involve the issue of restriction of range. Restriction of range means that, as a result of selection, the range of subjects in a study is inevitably restricted and only those who are selected with certain

standards rather than those who are randomly drawn from the true population are available for investigation. Restriction of range leads to a sampling bias. In TOEFL's prediction studies, the sample under study is restricted by the minimum TOEFL requirement for admission selection so that an unrandomized sampling bias occurs. Based on a critical analysis of six studies of TOEFL's predictive validity, Yan (1994) found that TOEFL means in these studies ranged from 491.00 to 561.00, which were above the 50th percentile rank in the population of all TOEFL takers. Standard deviations in these studies ranged from 38.80 to 66.00, which were lower than the standard deviation of the population. In most cases, sampling in the TOEFL/GPA studies was based primarily upon availability of subjects instead of randomization. This easily produces a biased sample with higher homogeneity than its population. A homogeneous sample will underestimate the predictive validity of TOEFL scores (Pedhazur, 1982; Cohen, 1983).

Predictor variables

There are different kinds of TOEFL scores used in TOEFL prediction studies. This variation in selection of predictor variables influences the estimation of TOEFL's predictive validity. Some examples are given as follows.

Firstly, some studies only used TOEFL total scores (Johnson, 1988; Light & Wan, 1991), some used the TOEFL sectional scores, others used total and sectional scores separately (Kwang & Dizney, 1970; Light, Xu & Mossop, 1987;

Hu, 1991) and a few used a combination of TOEFL total scores and sectional scores as one predictor. The predictive validity varies on the basis of single scores or composite scores of TOEFL.

Secondly, from 1963 to 1976 TOEFL consisted of five subtests. The five-section TOEFL had 200 total items and required two hours and 20 minutes of administration time. Some prediction studies examined the five-section TOEFL (Harcey, 1979; Bositc, 1981; Stover, 1982). The current three-section TOEFL consists of 150 items and requires one hour and 45 minutes of actual testing time. Some studies explored the predictive validity of the three-section TOEFL(Martin, 1971; Sharon, 1972; Shay, 1975; Riggs, 1982). Because of differences in section construction items included and time allocated for the two forms of TOEFL, special caution has to be taken when one compares the predictive validity of TOEFL scores obtained over time from different test administrations. However, this was unfortunately ignored in some TOEFL prediction studies (e.g., Odunze, 1982).

Thirdly, English language ability can be affected over a short period of time by additional training or lack of pre-test practice (ETS, 1994a). Thus ETS set a rule that a TOEFL score report will only be valid for two years. However, even within two years the range of time to take TOEFL is still important. The research literature documented the TOEFL prediction studies with a variety of timings, such

as summer TOEFL scores after arrival in the USA (Gue & Holdaway, 1973), pre-instruction TOEFL scores (Schrader & Pitcher, 1970), and pre-study TOEFL scores (Light & Wan, 1991). Most studies used pre-admission TOEFL scores (e.g., Heil, & Aleamoni, 1974; Ayers & Peters, 1977), except a study using after-admission TOEFL scores (Ho & Spinks, 1985). It is important to note that the time lapse between the collection of predictor scores and the collection of the criterion scores will impact the predictive validity of TOEFL scores. Furthermore, pre- and post-admission scores affect significantly the degree of homogeneity of the sample. The former will be much more heterogeneous, and the latter will result in a fairly selective sample.

Besides various forms of TOEFL scores, many TOEFL prediction studies used other language test scores, obtained from such standardized or local tests as Lado Test B and C (Chase & Stallings, 1966), the Pennstat (Chase & Stallings, 1966), Test of the American Language Institute at Georgetown University (AACRAO, 1971), Michigan Test of English language Proficiency (MTELP) (Pack, 1972; Abadzi, 1976), the English Placement Examination (Heil & Alaemini, 1974), the GRE general test's verb subtest (GRE-V) (Ayers & Peters, 1977), and Wechsler Adult Intelligence Scale's form R Vocabulary subtest (WAIS-R-V) (Hassen, 1982), as predictors. Other studies used writing scores and interview scores (Gue & Holdaway, 1973), ESL course average grade (Stover, 1982; South, 1992) as the predictor of academic success.

Quite a few studies have used non-language predictors, such as GRE-Q (Ayers & Peters, 1977; Yule & Hoffman, 1990; Ayers & Quattlebaum, 1992); high school GPA, age, years out of school, resident status, cultural background (Andalib, 1976); WAIR-R-V, SAT (Wilox, 1975); native language, major area of study (Stove, 1982); ratings of quality of academic performance (AACRAO, 1971); and LAST (Schrader & Pitcher, 1970).

Criterion variables

Selection of criterion variables is a crucial but difficult task in designing a prediction study. Although it has always been criticized (e.g., Graham, 1987), GPA is still the most frequently used criterion variable. This is largely because (a) it is the most typical (if not perfect) indicator of academic success (Wimberley, McCloud, & Flinn, 1992); (b) it is the most readily accessible criteria for academic achievement (Light, Xu & Mossop, 1987); and (c) it is relatively well-defined and widely understood (Young, 1993). However, different versions of GPA have been seen in the research literature.

Types of GPAs in terms of a period of time include:

- First-term GPA (Pack, 1972; Stove, 1982; Wilcox, 1975; Light & Wan, 1991; Light, Xu & Mossop, 1987; Kwang & Dizney, 1970);
- First- and second-term GPA (Abdzi, 1967; Harvey, 1979; Hell & Aleamoni, 1974; Martin, 1971; Odunze, 1982);

- First-year GPA (Chase and Stallins, 1966; Riggs, 1982; AACRAC, 1971; Gue & Holdaway, 1973; Schrader & Pitch, 1970);
- First-year, one-and-half-year, and two-year GPA (Yule and Hoffman, 1990);
- Graduation GPA (Ayers & Peters, 1977); Ayers & Quattlebaum, 1992);
- GPA obtained from unreported or unable to identified terms (Hassen, 1982; Andalib, 1976; Sharon, 1972; Hu, 1991; Johnson, 1988).

Types of GPAs in terms of different point systems include:

- Four-point GPA (AACRAO, 1971; Martin, 1971; Light & Wan, 1990; Ayers & Peters, 1977);
- Five-point GPA (Andalib, 1976);
- Nine-point GPA (Gue & Holdaway, 1973);
- Percentage GPA (UBC, 1993);
- Letter grade GPA (UBC, 1993).

Other criteria used include:

- Numbers of credit hours (Shay, 1975; Abdzi, 1967; Johnson, 1988);
- Average of 12 credits successful completed (Light & Wan, 1991);
- Verbal- and nonverbal-course GPA (Bostic, 1981);
- Academic index, advisor's rating (AACRAO, 1971);
- Eventual TA recombination (Yule & Hoffman, 1990);
- Average accumulated credit per semester

(Christopher, 1993).

Besides the above, as an index to academic achievement GPA also varies in sections, courses, instructors, majors, years, programs, and institutions, as well as countries. Various versions of GPA will have a significant impact on the estimation of TOEFL's predictive validity. Therefore, each TOEFL prediction study should specify and justify what kind of GPA is used.

Result interpretation

Various ways of interpreting results are another source of inconsistency in research findings of TOEFL's predictive validity. There were two consistent problems regarding result interpretation in the TOEFL prediction studies.

First, there have been neither consistent standards nor conventional terminology used to evaluate whether a measure of the TOEFL's predictive validity in a study is high or low. Some studies claimed that TOEFL scores were a useful, reliable, significant, meaningful, adequate, strong, and satisfactory predictor of GPA respectively (Chase & Stallings, 1966; ; Hwang & Dizney, 1970; Shay, 1975; Ayers & Peter, 1977; Sokari, 1981; Odunze, 1982; Ayers & Quattlebaum, 1992). On the contrary, other studies declared that TOEFL scores were of limited, doubtful, and questionable value in predicting GPA respectively (Harvey, 1979; Bostic, 1981). Obviously here, what was meant by useful, doubtful, or other modifiers was rather vague and subjective. As a matter of fact, what is deemed useful,

effective, or strong by one researcher may be deemed useless, ineffective, or weak by another researcher or by the same one at another context. For example, based on the research finding (r=.14, p<.05), one study (Light, Xu & Mossop, 1987) asserted that the correlation was too low to have any practical significance and therefore TOEFL was not an effective predictor of academic success. However, the researchers explained neither why an r of .14 was too low nor what standards were used to reject the TOEFL's predictive validity in the study. Thus, their conclusion is arbitrary.

To determine the strength, importance, and meaningfulness of findings, an estimate of effect size instead of testing of statistical significance is generally recommended (Cohen, 1988; Pedhazur & Schmelkin, 1991). Cohen (1988) proposed conventions for small, medium, and large effect sizes for correlation coefficients, regression coefficients, and differences between means. The results of TOEFL prediction studies should be interpreted according to well-established standards like Cohen's conventional definitions on R^2 (1988) to avoid subjectiveness and arbitrariness.

The second problem is about what estimates should be used to judge the relative importance among predictor variables. Some studies concluded that TOEFL scores were a better, higher, best, strongest, or lower predictor by comparing the scores with other predictors in terms of

correlation coefficients or regression coefficients obtained (Wilcox, 1975; Chase & Stallings; Ho & Spinks, 1985; AACRAO, 1971). These studies judged the predictors' relative importance on the basis of significance test results on improper estimates such as $\underline{r}$ or $\underline{R}^2$. Conventionally, change in $\underline{R}^2$ or squared semipartial correlation is recommended to estimate the relative importance among predictor variables (Pedhazur, 1982; Tabachnick & Fidell, 1989).

Summary

The literature review in this chapter helps to build both a conceptual and a methodological bases for estimating the predictive validity of TOEFL scores on GPA.

Theoretically, it was revealed that numerous factors in social, educational, linguistic, psychological, and physiological domains influence academic achievement. There is no one single factor which can fully determine academic success or failure. The unique contribution of any single factor to academic achievement should be examined in comparison with other relevant factors. Therefore, to investigate TOEFL's predictive validity, one should consider it within a comprehensive context which includes both language factors and non-language factors. For language factors, one should further consider different language proficiency (e.g., CALP, BICS), language skills (e.g., listening, writing), or linguistic registers in subject areas (e.g., mathematics, science).

From a methodological perspective, there are five major aspects which have significant effects on the estimation of TOEFL's predictive validity: analytical models, subject variables, predictors, criteria, and result interpretation. They deserve special attention in the research design in order to ensure satisfactory research validity.

Chapter Three

Method

This chapter outlines the method of the present study, including the program setting, participants, the predictor variables, the criterion variable, the analytical model, operational definitions of predictive validity, and research hypotheses.

The program setting

The UBC/Ritsumeikan Academic Exchange Program began in 1991 based upon an agreement for the establishment of an international academic exchange between the University of British Columbia and Ritsumeikan University. It is the largest exchange program of this type in North America. The program operates on an eight-month basis. Each year about 100 undergraduate students from Ritsumeikan University study at UBC from September to April as a part of their four-year university education. After that, they go back to continue their studies in Japan.

Ritsumeikan University was originally founded in 1869 by Japanese Prince Saionji Kinmochi and is one of the private universities in Japan. It presently comprises seven Colleges and seven Graduate Schools in Law, Economics, Business Administration, Social Sciences, International Relations, Letters, Science and Engineering. The total enrollment of students in the 1992-93 academic year was about 25,000, of which undergraduate students were over 23,000 (Ritsumeikan University, 1993). The ratio of success

in application for admission into the University is about 1:20.

Applicants to the UBC/Ritsumeikan Academic Exchange Program are required to submit their academic records, TOEFL official score reports, as well as writing samples in English for evaluation. To help applicants prepare for writing TOEFL, Ritsumeikan University provides TOEFL preparation workshops. Based upon both academic aptitude and English proficiency, Ritsumeikan University selects about 100 qualifiers into the program from the pool of applicants in second- and third-year courses.

The program provides a content-oriented curriculum with an emphasis on cross-cultural communication. The instructors are from the Department of Language Education at UBC. English is used as the medium of instruction. At the beginning of the program, the students are grouped into five classes. Each class included about 20 students with one teaching assistant. They are required to complete six three-credit courses in one academic year, three three-credit courses for each term. All of them take courses offered by the Department of Language Education in the first term. The courses offered in 1993-94 included: Intercultural Communication in Second Language Education, Communication Skills in Educational Settings, Academic Discourse in Second Language Education, and Second Language Education Practicum. In the second term, those whose TOEFL total scores meet the UBC minimum requirement of 550 may attend regular UBC

classes in the Faculty of Arts and other faculties for which they have pre-requisites. The credits the students earn at UBC are transferable to their home university.[4]

All the program students live in the UBC/Ritsumeikan House on the campus of UBC. Pairs share an apartment with two Canadian roommates. In addition to daily life experience, field studies, a buddy program, and other programs are arranged to enhance the students' cross-cultural understanding of Canadian society. The students are also involved in social activities on and off campus, such as a seminar series by the UBC Pacific Rim Club and volunteer work at preschools.

Participants

The target population of the study was the UBC/Ritsumeikan Academic Exchange Program students. The sample was a total of 97 students who enrolled in the 1993-1994 program. Among them, 46 students were male and 52 female. The range in age was from 20 to 23 years old, except one senior student aged over 60. They were all second year undergraduate students at Ritsumeikan University. Ninety five students majored in the humanities and social sciences such as Law, Business Administration, International Relation, Economics, and English Literature, and only two in Engineering. Japanese is their first language and most of

[4] According to William McMichael (W. McMichael. personal communication, April, 1995), current academic coordinator of the program, the 1994-95 program has adjusted its curriculum structure and the 1995-96 program will have larger changes.

them had not yet experienced studying and/or staying in North America before the program. Most were at the intermediate level in English proficiency. Their TOEFL total score mean was 515.96 with a standard deviation of 26.03.

It was evident that the sample was quite homogeneous in terms of age, native language, country of origin, cultural background, major fields, and English proficiency.

The predictor variables

Seven predictor variables were used in the study based upon suitability to the research question and availability in the UBC/Ritsumeikan Academic Exchange program. These predictor variables were: TOEFL total scores, TOEFL section I scores, TOEFL section II scores, TOEFL section III scores, oral interview scores, writing sample scores, and gender.

TOEFL scores, including total scores and three subscores, served as predictor variables in the study. These scores were obtained from the different TOEFL administrations at Ritsumeikan University through the Institution Testing Program (ITP) from January through May of 1993 when the students applied for admissions into the program.

Two things should be noted. First, the highest TOEFL score for each student was used in the study. Most students in the program wrote the TOEFL repeatedly in different administrations. There are three frequently seen alternatives for use of TOEFL scores for admissions: the highest TOEFL score, the latest TOEFL score, or the average

TOEFL score. Both Ritsumeikan University and UBC consistently used applicants' highest TOEFL scores to evaluate English language proficiency for program admission.

Second, TOEFL scores were obtained through the Institution Testing Program rather than the regular Friday and Saturday Testing Programs.[5] ETS states that TOEFL scores under the ITP are not acceptable for official admission purposes. However, the study had to use the ITP TOEFL scores because they were the only alternative Ritsumeikan University administered for the program applicants. UBC and Ritsumeikan University agreed to use ITP TOEFL scores for program admissions purposes. According to the TOEFL Test Manual (ETS, 1992), the ITP Manual (ETS, 1994f), and discussions (Kantor, R. N., personal e-mail communications, 1994 & 1995) between the author of the current thesis and Dr. Kantor, Director of TOEFL Program Office, the ITP TOEFL scores are still considered substantially valid and are comparable to scores earned under the regular programs.

Two other English proficiency measurements were available in the program and used as predictor variables. They were the September oral speaking scores and the September writing sample scores. The purpose of those two

---

[5] There are two different kinds of TOEFL testing programs according to the TOEFL Test and Scores Manual (ETS, 1992). The official TOEFL testing programs, including Friday and Saturday testing programs, are administrated internationally in the TOEFL testing centers. The Institution Testing Program, whose items were previously used in the official testing programs, is administrated at local institutes around the world.

measurements was to evaluate English speaking skill and writing skill respectively before instruction started, while TOEFL does not provide direct information on writing and speaking skills. The validity and reliability of these two measurements have not been reported.

The September oral proficiency interview took 20 minutes. Each student's oral performance was rated by the interviewers on a 0-5 11-point scale (including extra five **plus marks**. See Berwick & McMichael, 1993, p. 3 & Appendix C). The interviewers received pre-interview training in oral proficiency interview and rating procedure.

The September writing sample scores were given by trained raters based on a 1-6 6-point scale (See Berwick & McMichael, 1993, p. 2 & Appendix B). Each student was required to write an essay on designated topics within 90 minutes. The assessment of the writing samples followed that of the TOEFL Test of Written English (TWE).[6]

In addition to the foregoing language-based predictors, Gender was used as a non-language predictor in the study.

The criterion variable

The criterion variable in the study was the first-term GPA. It was calculated on the basis of a percentage grading and credit weighting system [7] which UBC adopted in 1991. At

---

[6] Both scores of the September oral proficiency interview and scores of the September writing sample will be labeled as speaking scores and writing scores respectively in the following text.

[7] The convertibility among different grading and credit weighting systems in North America is beyond the scope of

UBC, course weight is expressed in credits. In general one credit represents one hour of instruction or two to three hours of laboratory work per week throughout one term. Courses are normally graded on a percentage basis with a corresponding letter grade assigned (UBC, 1993).

The first term GPA included the average percentage grades in three courses. They were: EDUC395A, Second Language Education Practicum; EDUC490A, Regional Studies In Second Language Education; and ENED379, Crosscultural Studies in Second Language Education. Each was a three credit course. Instructors in the Department of Language Education taught the courses and assessed academic achievement. The course grades were given based upon a set of specific criteria outlined in the various course syllabi at the beginning of the term (Berwick & McMichael, 1992; Berwick & McMichael, 1993). Table 3.1 shows that the set of criteria mainly placed weights on written tasks to evaluate students' academic achievement.

Analytic model

This study used multiple regression analysis as the analytic model. The study focused on the estimate of the predictive validity of TOEFL scores on GPA. Hence the regression model is appropriate and intimately related to the primary goal of the study. Furthermore, the complexity of the research problem under study required a

---

the present study. For detailed discussion on this issue see Cohen & Cohen (1983) and Pedhazur (1982).

Table 3.1

Grade criteria on different aspects in the three courses

	EDUC395A (%)	EDUC490A (%)	ENED379 (%)
Field work journal	20	20	
Oral presentation	20	15	10
Term paper	20		20
Lab work			
Final Examinations		25	15
Assignments	30	30	35
Progress evaluations		30	
Bibliography	10		
Literature review			10
Participation			10
TOTAL	100	100	100

powerful analytic tool. As a highly general and very flexible data-analytic system (Cohen & Cohen, 1983), the regression model, particularly the multiple regression model, can be applied to investigate various factors related to the predictive power of the TOEFL score. The data of the study was processed with SPSS for Windows (Release 6.0).

Operational definitions of predictive validity.

The present study utilized change in squared R (ΔR^2) as the estimator to assess predictive validity. According to Cohen's conventional definitions (Cohen, 1988, pp. 412-414), .02, .13, and .26 are respectively defined as small, medium,

and large effect size of R^2. Based on Cohen's definitions, the present study defined four operational levels of predictive validity (see Table 3.2).

Table 3.2

Four levels of Predictive Validity

Level	R^2	Predictive validity
1.	.000 - .019	Negligible
2.	.020 - .129	Small
3.	.130 - .259	Medium
4.	.260 - 1.00	Large

Research Hypotheses

The present study advanced the following research hypotheses for testing:

1. TOEFL total scores have predictive validity on first term's GPA for the UBC/Ritsumeikan Exchange Program students.

2. TOEFL sectional scores have predictive validity on first term's GPA for the UBC/Ritsumeikan Exchange Program students.

3. Writing scores have predictive validity on first term's GPA for the UBC/Ritsumeikan Exchange Program students.

4. Speaking scores have predictive validity on first term's GPA for the UBC/Ritsumeikan Exchange Program students.

5. Gender has predictive validity on first term's GPA for the UBC/Ritsumeikan Exchange Program students.

Summary

This chapter delineated the research design of the present study. Participants were 97 Japanese exchange students. The study employed a multiple linear regression model to analyze the relationships of TOEFL scores and other predictor variables to first term's GPA. Four operational levels of predictive validity were defined for result interpretation. The study tested five research hypotheses.

Chapter Four

Results

This chapter summarizes treatment of the missing data, steps taken to check for violation of assumptions, and an analysis of the descriptive data. The chapter presents the main findings of a multiple regression analysis.

Treatment of the missing data

An examination of the data file used in the present study showed that there were three cases with missing values and one case which had a suspicious value on the TOEFL section II score.

Since only three missing-value cases were found from a sample of 97, there were very few chances that a systematic pattern existed among the missing-value cases. In other words, there were reasons to believe that the missing values for the variables occurred randomly. Therefore, the listwise missing-value treatment was employed in the study. This treatment keeps all variables but eliminates the missing-value cases. It is also the default for the missing-value treatment in the SPSS for Window program. Three cases, two with missing values in speaking scores and one in GPA, were eliminated from the data file.

For case 40, the TOEFL total score was 570, with three sectional scores 50, 68, and 53 respectively. The TOEFL section II score was suspicious. Note that 68 is the maximum score in Section II. It was almost impossible to reach it while the other sectional scores were around 50. It was also

found that in a TOEFL test administrated about two months earlier than the currently discussed test the same person scored only 513. It was unlikely that this student would gain about 60 points within two months. Thus the section II score of 68 might be a data entry error. Since all the original reports of TOEFL scores were at Ritsumeikan University in Japan, it was impossible to check this particular TOEFL score. Case 42 was, therefore, excluded from the data.

Descriptive statistical analysis

Means and standard deviations of all the variables are shown in Table 4.1 below. The mean TOEFL total score in the study was 515.96 and sectional scores were 49.74, 53.58, and 51.48 respectively. Standard deviation (SD) of the TOEFL total score was 26.03.

Table 4.1

Means and standard deviations of all the variables

	GPA	GENDER	SPEAK	WRITE	TOEFL	SEC1	SEC2	SEC3
N	93	93	93	93	93	93	93	93
M	71.97	1.55	1.46	2.71	515.96	49.74	53.58	51.48
SD	7.54	.50	.83	.83	26.03	4.16	3.34	2.91

ETS reported that based on the total of 1,338,682 examinees tested between July 1991 to June 1993, the mean

TOEFL total score was 519.00 and SD was 68.00. The mean TOEFL total scores and the mean sectional scores were 490.00, 49.00, 50.00, and 48.00 respectively for the group of the test takers whose native language is Japanese,(ETS, 1993).

Means of three groups (the total group, the group of Japanese examinees, and the 1993-94 program students) were similar (519.00, 490.00, and 515.96), but SDs of the sample under study were almost three times smaller than those of the total group. The considerable difference in SD between the total group and the sample under study indicates that the sample was homogeneous in terms of TOEFL scores. Obviously, this is because the sample was restricted to successful applicants whose TOEFL scores met the minimum TOEFL score, rather than to applicants randomly selected from the true population.

Table 4.2 below shows a Pearson Correlation matrix among the variables.

When the correlation matrix of variables is obtained, it is necessary to perform an omnibus test to make sure there is an overall significant interrelation existing among each pair of correlations in the matrix (Cohen & Cohen, 1983, p. 85 & pp. 315-316). If there is no overall significant relationship among the correlation, then the overall relationship in the matrix results from random sampling error rather than from the meaningful association

Table 4.2

Pearson correlation matrix of the variables

	GPA	GENDER	SPEAK	WRITE	SEC1	SEC2	SEC3	TOEFL
GPA	1.000	.457	.139	.342	.284	.332	.193	.365
	.	.000	.184	.001	.006	.001	.063	.000
GENDER		1.000	.003	.283	.252	.067	-.065	.139
		.	.979	.006	.015	.520	.537	.185
SPEAK			1.000	.358	.355	.238	.165	.352
			.	.000	.000	.022	.114	.001
WRITE				1.000	.183	.297	.068	.248
				.	.079	.004	.517	.016
SEC1					1.000	.313	.347	.794
					.	.002	.001	.000
SEC2						1.000	.368	.731
						.	.000	.000
SEC3							1.000	.713
							.	.000
TOEFL								1.000
								.

between each pair of variables. This is almost the same as performing an omnibus F-test before post hoc t-tests for means of each group in ANOVA.

In the present study, a Bartlett Chi-square test was performed to test the overall null hypothesis that all possible sample correlations among the set of variables in the matrix were zero. The result rejects the null hypothesis ($p<.001$). This indicates that there is a significant

interrelation among the entire set of Pearson correlation coefficients.

Checking for violation of assumptions

Each analytical model, such as the correlation model and the regression model, has been developed based on certain essential assumptions. Intelligent use of analytical models must meet the assumptions underlying the models. Violations of assumptions lead to estimate biases. Therefore, checking for violation of underlying assumptions is considered to be an indispensable component inherent in a regression analysis. The following sections will discuss (a) two general assumptions underlying any analytic model, i.e., the assumption of specification errors and the assumption of measurement errors; (b) six specific assumptions underlying a regression analytic model (Berry, 1993); (c) outliers and influential points.

The assumption of specification error. This assumption requires that an analytic model should flawlessly reflect its underlying rationale regarding the effect of independent variables on dependent variables. There are three types of specification errors: (a) omission of relevant variables into the regression model; (b) incorrect specification of the manner in which independent variables affect the dependent variables and (c) inclusion of irrelevant variables (Pedhazur & Schmelkin, 1991, pp. 389-390). Specification errors are the most damaging as they pose the most serious threat to valid interpretation of regression

results. However, it is difficult to tell whether all relevant variables have been included in the model, if all irrelevant variables have been excluded, or whether the model has been correctly specified in the context of social science research. The practical way to avoid specification errors is to use a well-grounded theory to build an analytic model. As Berry (1993) has pointed out, people should judge regression models by whether these models conform to their theories, and thus whether the models can be used to answer their research questions (P. 8).

To reduce specification errors in the present study, the following efforts were made within data and time constraints.

1. Variables were selected for a regression analysis based upon knowledge about language and non-language factors that influence academic achievement (see chapter two). The present study used language-based predictors and also introduced an exploratory non-language variable, gender, into the regression.

2. The study focused on accurately estimating the unique contribution of TOEFL scores on GPA, rather than on measuring the effects of all the variables in the regression model. This is because the primary research interest is to know how well TOEFL scores, as a single predictor, can predict GPA, not how much variance in GPA can be explained.

Assumption of measurement errors. This assumption assumes that all variables under study are measured without

error. In reality, test scores unavoidably include measurement error. Berry (1993) has provided an extensive discussion of three types of measurement errors: random measurement errors, non-random measurement errors, and measurement errors involving the use of proxy variables (pp. 49-60).

The present study dealt with the issue of measurement error in two ways:

1. Information about measurement of indicators, GPA, TOEFL scores, speaking scores, and writing scores, was gathered. It is almost impossible to perform a measurement without error in social science research. Information about all measures used in the study were gathered in order to identify possible measurement errors. In the present study, both GPA and TOEFL scores are among the most frequently used indicators in educational practice, although the quality of these measures have been long debated. Both speaking scores and writing scores are locally used within the UBC/Ritsumeikan Program. Thus there is sufficient information available about TOEFL and GPA, but not much about speaking scores and writing scores.

2. The findings were interpreted with special care. The regression model does not provide sufficient power to handle measurement errors like LISREL does. Therefore, the present study clearly distinguishes the difference between TOEFL and language proficiency, and between GPA and academic achievement. With these distinctions in mind, results of the

regression analysis were interpreted carefully so as to avoid overgeneralization.

The assumption of linearity. The nature of the relationships between predictors and criteria, linearity or non-linearity, requires a proper model for regression analysis. As shown in Figure 4.1, the residuals are randomly distributed and there are no systematic patterns existing between the predicted values and the residuals. This justifies the use of the linear regression model.

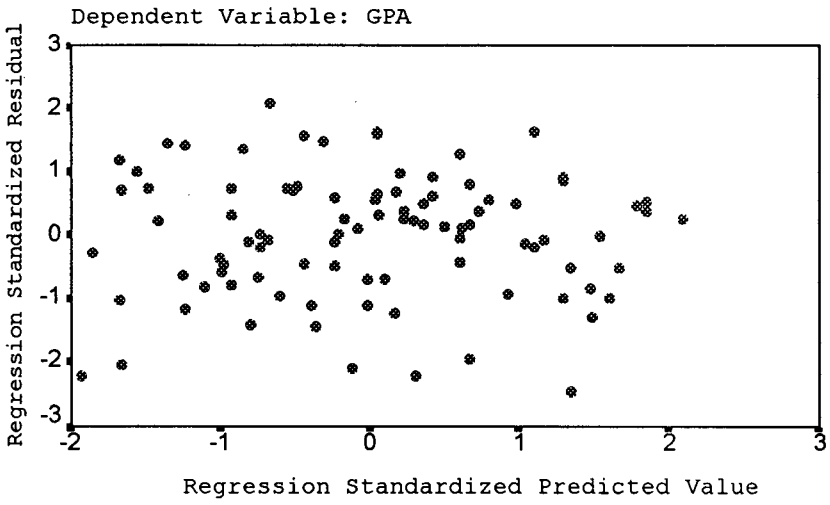

Figure 4.1. Scatterplot of the distribution of the residuals.

The assumption that mean of the residual is zero. This assumption means that the variance of the residuals is constant for all levels of the independent variables. Figure 4.1 also shows that the spread of the residuals does not increase or decrease with the magnitude of the predicted

values on the X axis. This indicates that the above assumption was met.

The assumption that residuals are independent. This assumption requires that residuals are independent of one another. Violation of this assumption, often referred to as autocorrelation, affects the validity of tests of significance. From Figure 4.1, we also can see that the residuals are randomly scattered above and below the zero horizontal band. This tells us that autocorrelation does not occur and the above assumption is met.

The assumption of normal distribution of residuals. This assumption requires that residuals should distribute normally. In the histogram of Figure 4.2, the distribution of the residuals appears approximately normal.

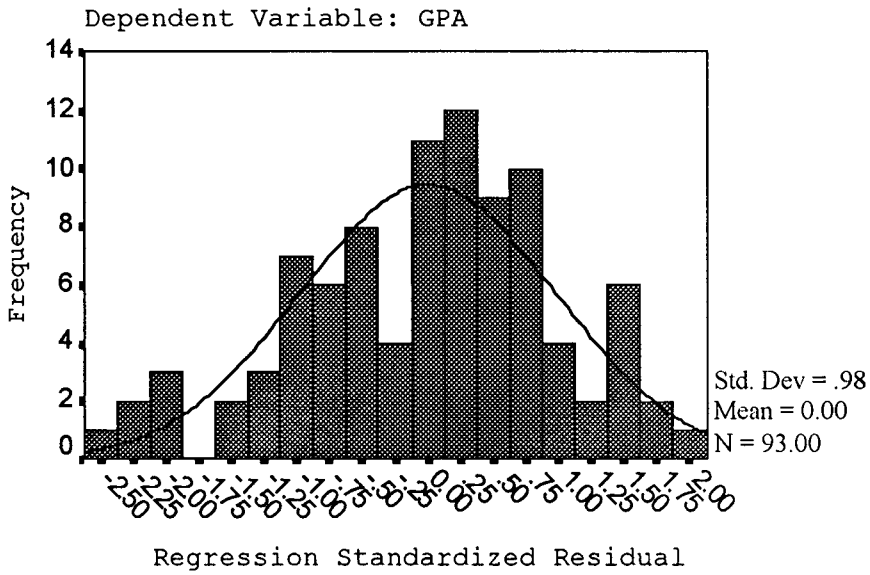

Figure 4.2. Distribution of residuals.

The assumption of the absence of perfect multicollinearity. This assumption assumes that there is no strong intercorrelations among independent variables. The existence of certain correlation among the independent variables indicates high multicollinearity. The tolerance of an independent variable is a commonly used measure of multicollinearity. In the present study, the tolerance of the predictor variables ranged from .63 to .89. This implies that the above assumption is satisfied.

The assumption that residuals are not correlated with each of the independent variables. In the present study, a correlation analysis was performed to check this assumption. The results showed that all the correlation coefficients between the independent variables and the residuals were .00 except that between gender and the residuals (r=.371).[8] Therefore, this assumption was also satisfied.

Diagnosis of outliers and influential points. Two frequently used measures, standardized residual and centered leverage were selected to detect outliers and influential points respectively. As shown in Appendix II, all the standardized residuals are below 3 units from zero, and thus no outliers are found. However, case 73 has a leverage of .241 which is twice as large as the upper limit of normal leverage values. It turned out that this case had a very low

[8] The considerable relationship between gender and the residuals clearly indicates again that gender is a composite variable which interrelates with other variables, known and unknown, or currently available and unavailable for research.

TOEFL total score, 487, but its GPA, 77, was five points higher than the GPA mean. Case 73, therefore, was identified as an influential point and eliminated before performing the multiple regression analysis.

Hierarchical regression analysis

The present study employs the multiple linear regression analysis with a hierarchical procedure instead of a stepwise procedure that is used most commonly. This decision is made based on the comparison among three options in the procedure of the multiple regression analysis.

The primary purpose of the present study was to estimate the predictive validity of TOEFL scores. That is, the study aimed at estimating the unique contribution of TOEFL scores to GPA, rather than the overall contribution of all predictors to GPA or the best linear combination of predictors to predict GPA, by partialling out the rest of the predictors under study. To accomplish this, the key issue was to determine the order or sequence of entering the predictors because different entry orders yield different estimates of the unique contribution of a predictor.

Generally speaking, there are three alternative procedures for the multiple regression analysis: simultaneous, stepwise, and hierarchical. In the simultaneous analysis, every predictor is entered into the regression analysis simultaneously and is partialled out from every other predictor indiscriminately. This procedure can provide a regression equation and squared R for all

predictors in the equation, but it does not estimate the unique contributions of each variable to the total variance in the equation.

Stepwise analysis can estimate the unique contribution of predictors by obtaining partial correlation or incremental variance. However, this procedure solely relies on statistical criteria to determine the sequence of entering predictors. When the competing predictors substantially connect with each other, the partial correlation or incremental variance might vary significantly according to the sequence in which predictors are entered. Thus, the procedure of stepwise analysis might create difficulties in estimating, interpreting, comparing, and replicating the regression results.

The hierarchical procedure enters predictors in a pre-specified sequence to estimate the unique contribution of each predictor to the total variance in the regression equation. The choice of a particular sequence of predictors is made in advance by the purpose and logic of the research, in contrast to the stepwise regression. The hierarchical procedure leads to tests of the hypotheses that define the order and improve our understanding of the phenomena under study (Cohen, 1983, pp. 120-125). As Tabachnick and Fidell (1989) discussed, simultaneous, stepwise, and hierarchical regression can be best used for model-estimating, model-building, and model-testing respectively (P. 150). Thus, in

the present study, hierarchical procedure was selected to test the hypotheses of the present study.

The sequence of entering the predictors in the present study was TOEFL total scores, writing scores, speaking scores, and gender. This sequence was mainly based on the research priority of the study because no causal relationship among the predictors was found. Since TOEFL scores reflected the major goal of the research and were the primary focus of the study, they were entered into the equation first. Writing scores and speaking scores followed because they were viewed as having lesser relevance to the research than TOEFL scores. Gender was entered last because it was used as an explanatory variable to exemplify non-language predictors' predictive validity which was not much documented in research literature. Within the TOEFL sectional scores, the sequence for entering was from sectional II, III, and I. This was based on a descending order in terms of three scores' predictive validity reported in the previous research (Abdzi, 1967; Aleamoni, 1974; Harvey, 1979; Heil & Johnson, 1988; Zirpoli, 1988).

In the following section, the first hierarchical regression analysis was performed to test the primary hypothesis regarding the predictive validity of TOEFL total scores. The second hierarchical regression analysis then mainly served to examine the unique contribution of each TOEFL sectional scores to GPA. Note that change in squared R

(ΔR^2) was used in SPSS as the estimator for unique contribution of each predictor to GPA in a hierarchical regression analysis. [9]

Hierarchical analysis with TOEFL total scores. Table 4.3 shows that the results of the hierarchical analysis with TOEFL total scores.

Table 4.3

Summary table of the hierarchical analysis with TOEFL total scores

Step	R	AdjR^2	F	P	ΔR^2	ΔF	ΔP	Variable
1	.377	.132	14.895	.000	.142	14.895	.000	In: TOEFL
2	.448	.183	11.184	.000	.059	6.555	.012	In: WRITE
3	.457	.182	7.731	.000	.008	.859	.357	In: SPEAK
4	.584	.311	11.251	.000	.132	14.470	.000	In: GENDER

In step 1, TOEFL total scores entered into the regression equation in order to determine the extent to

[9] In the present study ΔR^2 is interpreted as the amount of variance added to R^2 by each predictor at the point that it enters the equation in a hierarchical procedure. For the distinguished differences in the meaning of unique contribution of a predictor to R^2 among the three procedures due to the differences in handling the overlapping among correlated predictors, see Tabachnick and Fidell, 1989, pp. 141-142 & pp. 150-154.

which the TOEFL overall score predicted GPA. Results showed that 14.20% of the variance in GPA was accounted for by the TOEFL total score. The rest of about 86% of the variance remained as residual or unexplained error which should not be misinterpreted as measurement error. This mainly implies that the amount of variance had not yet been explained.

In step 2, writing scores entered into the regression equation. The result showed that it accounted for 5.90% of the variance. By adding writing scores, the squared R in the equation increased to .183.

In step 3, the addition of speaking scores to the regression equation only increased 0.70% of variance accounted for. This may indicate that speaking proficiency of the students did not contribute to their academic achievement significantly.

Step 4 was used to examine the effect of the addition of an explanatory non-language variable to the regression model. By addition of gender, the squared R increased to .31 and the squared R change was .132. This showed that gender difference placed one of the largest weight on GPA.

To further analyze the effect of gender difference, an ANOVA on TOEFL total scores and gender was performed. Results showed that there was no significant difference of TOEFL scores due to gender, F (1,90) = 1.476, $p > .05$. This implied that the variance of GPA was not due to gender difference in TOEFL total scores. In other words, the difference in academic achievement appeared not to be

affected significantly by the different levels of English proficiency, but rather, by the differences in non-language factors among male and female students.

Hierarchical analysis with TOEFL sectional scores. Table 4.4 shows that the results of a hierarchical analysis with TOEFL sectional scores.

Table 4.4

Summary table of the hierarchical analysis with TOEFL sectional scores

Step	R	AdjR^2	F	P	ΔR^2	ΔF	ΔP	Variable
1	.334	.102	11.297	.001	.112	11.297	.001	In: SEC2
2	.341	.096	5.857	.004	.005	.482	.489	In: SEC3
3	.400	.131	5.592	.002	.044	4.590	.035	In: SEC1
4	.455	.170	5.674	.000	.047	5.130	.026	In: WRITE
5	.466	.171	4.762	.007	.010	1.093	.299	In: SPEAK
6	.589	.300	7.513	.000	.130	16.870	.000	In: GENDER

The second hierarchical analysis used three TOEFL subscores as the predictors rather than a composite TOEFL score. Results showed that, while high tolerances for three sectional scores revealed low interrelations among the subscores, TOEFL section II had the highest squared R

change(ΔR^2 =.112), compared with section I (ΔR^2 =.044), section III (ΔR^2 =.005). In other words, section II scores among them had the most importance impact on the variance in GPA.

The unique contribution of writing scores to the equation was .047, speaking scores was .010, gender was .130. Compared with writing scores .059, speaking scores .008, and gender .132 as shown in the first hierarchical analysis, the results indicated that the pattern of the relative importance among the three predictors did not change, while TOEFL total scores were partitioned into three sectional scores.

Summary

The results of the present study show that the predictive validity of TOEFL total scores was .142 ($p<.001$). For TOEFL sectional scores, section II scores were the most important of three sectional scores. Among all predictors in the study, gender had the highest predictive validity on GPA when TOEFL sectional scores were used. Gender, TOEFL total scores, writing scores, speaking scores accounted for 31.00% ($p<.001$) of the variance in GPA. Gender, TOEFL sectional scores, writing scores, speaking scores accounted for 30.04% of the variance in GPA ($p<.001$).

Chapter Five

Discussion

This chapter discusses findings pertinent to the research hypotheses, interprets the implications of these findings, and draws conclusions of the study.

Predictive validity of TOEFL total scores on GPA

The present study examined the predictive validity of a predictor variable in two ways: (a) evaluation of its unique contribution to GPA on the basis of the four operational levels of predictive validity (see Chapter three) and (b) assessment of its relative importance in comparison with other predictors under study.

Results of the present study show that ΔR^2 of TOEFL total scores is .142 ($p<.001$). This result is comparable with results of a meta-analysis of 27 TOEFL prediction studies in which the mean correlation coefficients of TOEFL total scores and first year's GPA is .300, i.e., R^2 is 9% (Yan, 1994). According to the operational levels of predictive validity, therefore, TOEFL total scores have a medium level of the predictive validity on first term's GPA.

The results also reveal that TOEFL total scores are ranked as the second largest among all the predictors under study and as the largest compared with the other two language-based predictors, writing and speaking scores. Thus, TOEFL total scores are an important predictor of GPA in the present study.

Based on these two findings reported above, it can be concluded that hypothesis I of the study is supported. That is, TOEFL total scores have a medium level of the predictive validity on first term's GPA for the group of students under study.

It is not surprising that TOEFL total scores only account for 14.20% of the variance in GPA. As shown in many studies, English language proficiency is just one of many factors affecting academic achievement. It appears that no single factor alone can completely or largely determine academic achievement. As the unique contribution of a single predictor to GPA, therefore, 14.20% clearly indicates that TOEFL total scores alone do explain a significant amount of variance in GPA. In other words, language proficiency by itself, among many other factors, does have an important effect on academic achievement.

We can further analyze how TOEFL's predictive validity on GPA is affected by each pair of TOEFL scores and GPA for each student under study.

As shown in Figure 5.1, we can divide each GPA-TOEFL pair into four divisions by using the mean GPA and the mean TOEFL total scores: upper left, upper right, lower left, and lower right. Both upper right and lower left divisions share one commonality: A high TOEFL score corresponds to a high GPA, or a low TOEFL score with a low GPA. However, both the upper left and lower right divisions show that a high TOEFL

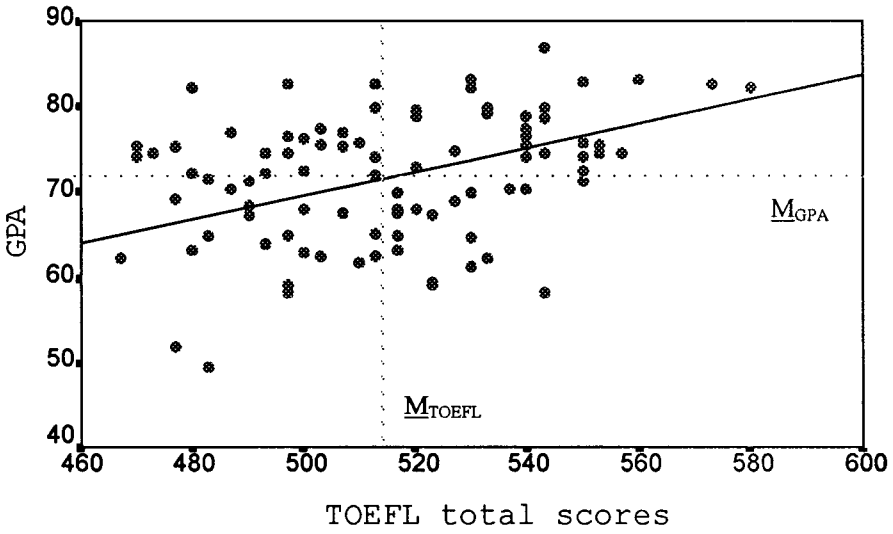

Figure 5.1 Scatterplot of TOEFL total scores and GRE

score goes with a low GPA, or a low TOEFL with a high GPA. Note that there are 24 cases in the upper left division, whereas only 15 in the lower right. Among these cases, there are at least 6 cases with TOEFL scores below 480 but their GPAs are above the mean GPA, whereas there is only one case with a TOEFL score above 540 and a GPA below 60. This indicates that in the present study about one quarter of the students who had low levels of language proficiency managed to achieve academic success. The number of this sub-group is higher than that of students who have the high level of language proficiency but are unable to reach the high level of academic achievement. In other words, a good TOEFL score does not necessarily guarantee a good GPA, but a low TOEFL score is often associated with a good GPA. It is these cases that might considerably decrease the magnitude of the

predictive validity of TOEFL scores. They prove again that for those whose native language is not English, many factors are involved in their academic learning at universities and language proficiency does not always function predominantly as a key element.

Predictive validity of TOEFL sectional scores on GPA

Results of the study show that the combination of three sectional scores accounts for 16.10% of the variance in GPA. This is close to what TOEFL total scores do. However, the unique contribution and relative importance of three sectional scores to GPA are remarkably different. It is shown that the changes in R^2 of section I, II, and III are .044 ($p>.05$), .112 ($P<.001$), and .005 ($p>.05$) which indicate they have small, medium, and negligible levels of predictive validity respectively. Section II scores have the second highest predictive validity among the six predictors under study and the highest among three TOEFL sectional scores. This finding is comparable with those in previous research (Johnson, 1988; Zirpoli, 1988; Light, Xu, & Morris, 1989). Thus, based on the uneven contributions of three sectional scores to GPA as well as their different importance, the conclusions for hypothesis II are: Section II have a medium level of predictive validity, section I scores have a small level of predictive validity, and section III scores have a negligible level of predictive validity.

There is an interesting question to ask: Among three sectional scores of TOEFL, why do section II scores tend to be so dominant in predicting GPA?

In the TOEFL test, section I, Listening Comprehension, measures the ability to understanding oral English; Section II, Structure and Written Expression, tests recognition of selected structural and grammatical knowledge in standard written English; And section III, Vocabulary and Reading Comprehension, tests the ability to understand written English (ETS, 1992, pp. 6-7). The three sections measure listening skills, writing knowledge,[10] and reading skills respectively. Thus, the findings presented here might indicate that a good GPA may demand more written English skills than spoken English skills (i.e., section II scores vs. section I scores). Furthermore, for written English skills, a good GPA may require more productive skills than receptive skills of written English (i.e., section II scores vs. section III scores). As seen in Table 3.1, about 85% of the course grades require written productive skills to fulfill various academic tasks such as term paper, course assignments, and final examinations. When students' academic achievement is assessed mainly based on performance in written expression, the weight of section II scores on GPA is greater than the other two sectional scores.

---

[10] ETS has not explained what is exactly meant by structure and written expression. Since section II uses both sentence correction and sentence completion to test basic knowledge about written English, the present study simply labels section II as writing knowledge instead of writing skills.

Predictive validity of writing scores on GPA

Results of the study reveal that the change of R^2 of writing scores is .059 ($p<.05$) when TOEFL total scores are used. This means that writing scores have a small level of predictive validity on GPA. It is also shown that the relative importance of writing scores are ranked third among four predictors, behind gender and TOEFL total scores, and before speaking scores. These findings to some extent support hypothesis III in the study and indicate that writing scores have a small level of predictive validity on GPA.

It is interesting to note that predictive validity of writing scores is substantially lower than that of section II scores (.047 vs. .112 for ΔR^2) when TOEFL sectional scores are used. Both deal with measurement of written English, but why do writing scores have so little contribution to GPA compared with its counterpart?

There might be tentative explanations to this question. For instance, as a locally used testing instrument, the writing sample assessment might not possess sufficient reliability and validity in measuring English writing skills as it should. This may result in under-estimation of its predictive validity on GPA. Also, Section II scores measure writing knowledge, while writing scores directly assess writing skills. For this group of Japanese students whose English is at the intermediate level, they might need more

basic writing knowledge of written English in order to fulfill their academic learning tasks successfully.

Predictive validity of speaking scores on GPA

The results of the study indicate that speaking scores have a negligible level of predictive validity on GPA (ΔR^2=.008, p>.05 or ΔR^2=.010, p>.05, depending on TOEFL total scores or sectional scores are used) and is consistently ranked as the least important among all the predictors in predicting GPA. Therefore, hypothesis IV regarding the predictive validity of speaking scores is rejected. Due to insufficiency of information about the reliability and validity of the oral interview used in this academic exchange program, the question concerning why speaking scores have a medium level of predictive validity on GPA must be left for future analysis.

Writing scores, speaking scores, and TOEFL scores in section I, II, and III could be seen to assess four language skills: listening, speaking, reading, and writing. It is interesting to look at weights of individual skills in predicting GPA as well as their overall effects on GPA.

First, the results of the study suggest that written skills in English are more important than oral skills in predicting GPA, as we have seen that ΔR^2 of section II scores is larger than that of section I and ΔR^2 of writing scores is larger than that of speaking scores. Also, the results tend to indicate that productive skills in written

English are more important than receptive skills since section II scores and writing scores have more predictive power on GPA than section III scores. However, we are unable to interpret the relative importance on GPA of written comprehension and aural comprehension, although ΔR^2 of section I scores exceed that of section III. All in all, it seems premature to draw a conclusion on the basis of findings of the present study about the relationship of those four language skills and academic achievement.

Second, results show that the cumulative R^2 of five language-based predictors, speaking scores, writing scores, and the three sectional scores, is .217 ($p<.001$). Thus, it could be inferred that the overall predictive validity of language factors on GPA might have an upper limit. Probably R^2 of language factors in a regression model would probably not exceed .25. In other words, among many other variables, language factors alone might optimize their contribution at about one-quarter of academic achievement assessed by GPA.

Predictive validity of gender on GPA

The results show that gender's ΔR^2 is .132, ($p<.001$) when TOEFL total scores are used, and .130 ($p<.001$) when the TOEFL sectional scores are used. The findings indicate that gender has a medium level of predictive validity on GPA. Gender is consistently ranked as one of the most powerful predictors under study. Therefore, it can be concluded that

gender is a good predictor in predicting GPA and hypothesis V is strongly supported.

Many studies have already proved that gender differences do influence academic achievement. However, it is still surprising that gender had such a remarkable contribution to GPA in the present study. Japanese female students as a group performed significantly better than male students in the course grades. This unanticipated finding raises a question: Why do gender differences affect GPA so greatly? In other words, why do female students academically excel over their male counterparts?

As reviewed in chapter two, research on gender differences reveals that gender is a composite factor influenced by and impacting on various factors in social, educational, linguistic, psychological, and physiological domains. Generally speaking, female students perform better in language arts and male students perform better in science. To find the possible cause for the gender difference in GPA, an F-test on gender difference in TOEFL scores was conducted. The results showed that there were no significant gender differences in these scores, although all three courses from which GPA were obtained were about language education and required good language proficiency. This result clearly indicates that gender differences in GPA are not caused by language factors but by other non-language factors. Probably non-language factors such as learning motivation, time spent on learning, academic aptitude,

learning style, previous knowledge, and cultural adaptability, might indirectly place effects on GPA through gender differences. Due to lack of data to analyze, what kinds of non-language factors and how they contribute to GPA for this group of students remain open for future research.

Implications

The findings in the present study may have practical implications for the UBC/Ritsumeikan Academic Exchange Program.

1. The main findings in the study clearly indicate that TOEFL total scores are a good predictor of first term's GPA for the UBC/Ritsumeikan Program students. Therefore, the program should continue to use TOEFL to measure English language proficiency for program admissions. Since TOEFL section II scores have the highest predictive validity among three sectional scores, they deserve particular attention for admission selection.

2. The findings on gender differences in GPA strongly suggest that non-language factors play an important role in academic achievement.[11] Thus, it is advisable that the

[11] This finding might lead mistakenly to another implication for program admissions: including more female students into the program and excluding more male ones from the program. In fact, this policy, given it was taken, would be not only politically incorrect but also logically oversimplified. As discussed in the previous chapters, the true reasons for gender difference in academic achievement are not due to sex difference, but rather, a combination of physical, cognitive, emotional, social factors embedded in gender difference. Therefore, for an intelligent educator, he or she should always find specific factors behind gender difference in academic achievement in order to help

program should gather as much information, particularly non-language data, as possible in order to select the most promising applicants. These types of information include previous GPA at Ritsumeikan University, letters of recommendation, scores in academic aptitude test, and personal statements of interests. More factors such as cultural knowledge, L1 level, motivation, intelligence, and personality, should be taken into consideration in making admission decisions.

3. The findings reveal that TOEFL scores alone do not absolutely ensure academic success. Thus, it is recommended that the currently used minimum TOEFL score of 550 not be used as a requirement for registration in regular UBC courses. Rather, an appropriate critical range of TOEFL scores should be established for program admission and management. In particular, for those who have low TOEFL scores but clearly show academic potential, the decision makers in the program should have a special policy for them so as to satisfy their learning needs and academic capabilities.

The present study may have theoretical implications. The issue of whether or not TOEFL scores can predict GPA has been debated for over 30 years. The study examined thoroughly the underlying rationale for TOEFL prediction studies and proposed a comprehensive framework for the

---

students, no matter male or female, to achieve their academic potentials.

analysis of factors affecting academic achievement. For its research design, the present study carefully considered technical treatments in predictor collection, criterion selection, analytical models, regression procedure, and validity levels in order to ensure the correct estimation of TOEFL scores' predictive validity. For these reasons, it may be thought that the present study might have taken a further step in solving the 30-year's TOEFL-GPA puzzle in terms of its comprehensive rationale and its improved methodology.

Limitations of the study

1. The present investigation did not include more relevant non-language predictors into the multiple regression analysis due to their current unavailability. This might cause possible specification errors in the multiple regression model used.

2. GPA used in the study was from three "Bridge Courses" designed specifically for the program. Compared to regular UBC courses, these courses may have different features such as course grade standards, instructor and teaching assistant's allocation, and communicative language environments. The uniqueness of this type of GPA might make uncertain the validity and generalizability of the study.

3. The study did not estimate the effect of restriction of range in TOEFL scores on the results of the multiple regressions analysis. Research literature has indicated that restriction of range in admissions will result in underestimation of the predictive validity, but we still

need empirical evidence to know to what extent and under what circumstances this underestimation may occur.

Directions for future research

1. The present study can be expanded into a time-series research project. On the basis of the data available for four years (1991-1995), we can examine the change pattern of TOEFL's predictive validity on first term's GPA over years. It is also feasible to compare the relationship of TOEFL's predictive validity to different kinds of GPA, such as second term's GPA and first year's GPA.

2. Other analytical models and statistical techniques can be used in the study. For instance, Multivariate Analysis of Variance can be used to analyze different GPA subscores; Hierarchical Linear Model can be adopted to examine the specific effects of different units such as individual, group, course, and instructor, on TOEFL's predictive validity; Linear Structural Relations can be employed to distinguish direct and indirect relationships among variables and assess the extent of measurement error that may appear.

3. A series of prediction studies can be developed to compare the predictive validity of TOEFL with those of other language tests such as Michigan Test of English language Proficiency (MTELP) and Certificate of Proficiency in English (CPE), and those of with aptitude tests such as Graduate Record Examinations (GRE) and Scholastic Aptitude Test (SAT).

4. It should be further examined how language proficiency, in particular, speaking, listening, reading, and writing, are related to academic achievement.

5. Case studies can be conducted to explore some special issues in depth. For instance, why are some students with good TOEFL scores unable to achieve academic success? Why do some other students eventually overcome their language problems and meet their academic challenges? What differences exist between female and male students in motivation, cultural adaptability, IQ, previous GPA, and other domains.

6. Decision theory (see Cronbach & Glaser, 1965) should be introduced in order to use TOEFL scores properly for admissions decision-making and program management.

Conclusion

The present study employed TOEFL scores as well as other predictors to predict first term's GPA with a multiple regression hierarchical analytic approach. For the UBC/Ritsumeikan Academic Exchange Program students, the following conclusions can be drawn from the findings of the present study:

1. TOEFL total scores alone have a medium level of predictive validity on first term's GPA.

2. TOEFL Section scores II, section I scores, and section III scores have the predictive validity on first term's GPA at medium, small, and negligible levels respectively.

3. Writing scores alone have a small level of predictive validity on first term's GPA.

4. Speaking scores alone have a negligible level of predictive validity on first term's GPA.

5. Gender alone has a medium level of predictive validity on first term's GPA.

Bibliography [12]

Adamson, H. D. (1990). ESL students' use of academic skills in content courses. English for Specific Purpose, 9, 67-87.

Alderman, D. L. (1982). Language Proficiency as a Moderator Variable in Testing Academic Aptitude. Journal of Educational Psychology 74, 580-87.

American Psychology Association. (1994). Publication Manual of American Psychology Association (4th ed.). Washington, DC: American Psychology Association.

American Psychology Association. (1985). Standards for educational and psychological tests. Washington, DC: American Psychology Association.

Ames, R. & Ames, C. (1984) Research on motivation in education (volume 1). Orlando, Florida: Academic Press, Inc.

Arena, L. (ed.) (1990). Language proficiency: defining, teaching, & testing. New York: Plenum Press.

Ayers, J. B. & Peters R. M. (1977). Predictive validity of the Test of English as a Foreign Language for Asian graduate students in engineering, chemistry, or mathematics. Educational & Psychological Measurement, 37, 461-463.

Ayers, J. B. & Quattlebaum, R. F. (1992). TOEFL Performance and Success in a Masters Program in Engineering. Educational and Psychological Measurement, 52, 973-75.

[12] The style of the present thesis followed consistently the four edition of Publication Manual of the American Psychological Association (APA, 1994) throughout the thesis.

Bachman, L. F. (1991). What does language testing have to offer? TESOL Quarterly, 25, 671-704.

Bachman, L. F. (1990). Fundamental considerations in language testing. Hong Kong: Oxford University Press.

Bachman, L. F., Davidson, F., & Foulkes J. (1990). A comparison of the abilities measured by the Cambridge and educational testing service EFL test batteries. Issues in Applied Linguistics, 1. 30-54.

Berry, W. (1993). Understanding regression assumptions. Beverly Hills, CA: Sage Publications, Inc.

Berry, W & Feldman, S. (1985). Multiple regression in practice. Beverly Hills, CA: Sage Publications, Inc.

Berwick, R. & McMichael, W. (1992). 1991-92 Ritsumeikan evaluation report. Unpublished manuscript, University of British Colombia, Vancouver, British Columbia, Canada.

Berwick, R. & McMichael, W. (1993). 1992-93 Ritsumeikan evaluation report. Unpublished manuscript, University of British Columbia, Vancouver, British Columbia, Canada.

Best, J. & Kahn, J. (1989). Research in education (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Black, J. (1991). Performance in English skills courses and overall academic achievement. TESL Canada Journal, 9, 42-56.

Bloom, B. et al. (1964). Taxonomy of educational objectives: Handbook II Affective domain. New York: David McKey Company, Inc.

Bloom, B. et al. (1956). Taxonomy of educational objectives: Handbook I Cognitive domain. New York: David McKey Company, Inc.

Bosher, S. & Rowekamp, J. (1992). Language proficiency and academic success: the refugee/immigrant in higher education. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED353914)

Bruner, J. (1966). Toward a theory of instruction. Cambridge, MA: Harvard University Press.

Buteyn, R. J. (1989). Gender and academic achievement in education. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED313103)

Carlson, S. B. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED268135)

Ching, R. J. & Moore, C. A. (1993). ESL assessment: What we learn when we open Pandora's box. Metropolitan Universities, 3, 35-46.

Choy, S. C., Davenport, B. M. (1986). The TOEFL: incomplete test of English proficiency. College Teaching, 34. 108-110.

Christopher, V. (1993). Direct and indirect placement test scores as measures of language proficiency and

predictors of academic success for ESL students. Unpublished master's thesis, University of British Columbia, Vancouver, British Columbia, Canada.

Cocking, R & Messier, J. (1988). Linguistic and cultural influences on learning mathematics. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: LEA.

Cohen, J & Cohen, P. (1983). Applied Multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: LEA.

Collier. V. P. (1989). How long? A synthesis of research on academic achievement in a second language. TESOL Quarterly, 23, 509-531.

Crandall, J. & Dale, T. (1987). ESL through content-area instruction: mathematics, science, social studies. Eaglewood Cliffs, NJ: Prentice-Hall.

Crowhurst, M. (1994). Language and learning across the curriculum. Scarborough, Ontario: Allyn & Bacon Canada.

Cummins, J. (1992). Language proficiency, bilingualism, and academic achievement. In Richard-Clmato, P. A. & Snow, M. A. (eds.) The multicultural classroom (pp. 16-69). London: Longman.

Cummins, J. (1991). Language development and academic learning. In Malave, L & Duguette, G. (Eds.). Language, Culture and cognition. (pp. 161-189). Clevedon, England: Multilingual Matters Ltd.

Cummins, J. (1984). Bilingualism and special education: issues in assessment and pedagogy. Clevedon, England: Multilingual Matters Ltd.

DeMauro, G. (1992). Examination of the Relationships among TSE, TWE, and TOEFL Scores. Language Testing, 9. 149-161.

Educational Testing Service. (1994a). TOEFL test and score manual supplement. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1994b). TOEFL Update. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1994c). 1994-95 Bulletin of information for TOEFL, TWE, and TSE. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1994d). The Researcher. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1994e). Checklist for using TOEFL scores. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1994f). TOEFL 1994-1995: Test of English as a foreign language Institutional testing program oversea edition. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1992a). Guidelines for the use of TOEFL scores. Princeton NJ: Educational Testing Service.

Educational Testing Service. (1992b). TOEFL test and score manual (1992-93). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1991). Guidelines for TOEFL institutional validity studies. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1989). The uses of standardized tests in American education. Princeton, NJ: Educational Testing Service.

Fischer, K. W. & Lazerson, A. (1984). Human development. New York: W. H. Freeman & Company.

Fletcher, J. & Stern, R. (1986). Language skills and adaptation: A study of foreign students in a Canadian university. Curriculum Inquiry, 19, 293-308.

Gradamn, H. L. & Hanania, E. (1991). Language learning background factors and ESL proficiency. Modern Language Journal, 75, 39-51.

Graham, J. G. (1987). English language proficiency and the prediction of academic success. TESOL Quarterly, 21, 505-521.

Gue, L. R. & Holdaway, E. A. (1973). English proficiency tests as predictors of success in graduate studies in education. Language Learning, 23, 89-103.

Hackett, G. et al. (1992). Gender, ethnicity, and social cognitive factors predicting the academic achievement of students in engineering. East Lansing, MI: National

Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. EJ454109)

Hale, G. A., et al. (1983). Effects of test disclosure on performance on the Test of English as a Foreign Language. Language Learning, 33. 449-464.

Hale, G.A. et al. (1984). A Comprehensive TOEFL Bibliography, 1963-82. Modern Language Journal, 68, 45-51.

Hale, G. A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. Language Testing, 5, 49-61.

Hale, G. A., Stanfield, C., & Duran, R. (1984). Summaries of Studies Involving the Test of English as a Foreign Language 1963-1982. Princeton, NJ: Educational Testing Service.

Ho, D. Y. F. & Spinks, J. A. (1985). Multivariate prediction of academic performance by Hong Kong University students. Contemporary Educational Psychology, 19, 249-259.

Hosley, D. Meredith, K. (1979). Inter- and intra-test correlates of the TOEFL. TESOL Quarterly, 13, 209-217.

Hu, S. P. (1991). English proficiency and academic performance of international graduate students. Dissertation Abstract International, 52, 1626-A.

Hughes, A. (1989). Testing for language teachers. New York: Cambridge University Press.

Husen T. & Postlethwaite, N. (Eds.)(1992). The international encyclopedia of educational (Vols. 1-12). New York: Pergamon.

Kishor, N. (1994). EPSE596 course notes. Unpublished manuscript, University of British Columbia, Vancouver, British Columbia, Canada.

Kwang, K. & Dizney, H. F. (1970). Predictive validity of the TOEFL for Chinese graduate students at an American university. Educational and Psychological Measurement, 30, 475-477.

Light, R. L., & Wan, T. (1991). Soviet students at U. S. Colleges: social perceptions, language proficiency, and academic achievement. TESOL Quarterly, 25, 179-185.

Light, R. L., Xu, M., & Mossop, J. (1987). English Proficiency and Academic Performance of International Students. TESOL Quarterly, 21, 251-61.

Johns A. M. (1981). Necessary English: a faculty survey. TESOL Quarterly, 15, 51-57.

Maccoby. E. & Jacklin C. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

Mestre, J. P. (1981). Predicting academic achievement among bilingual Hispanic college technical students, Educational and Psychological Measurement, 41, 1266-1264.

Mohan, B. (1986). Language and content. Reading, Mass: Addison-Wesley.

Morgen, B. S. (1990). A comparative study of the use of standardized English language proficiency tests by U. S. graduate schools. College and University, 65, 295-307.

Mullein, K. A. 1978). Direct evaluation of second language proficiency: the effect of rater and scale in oral interviews. Language Learning, 28, 301-308.

Norusis, M. (1994). SPSS-Window: introductory Statistics Guide (for SPSS-window release 6.0). Chicago, IL: SPSS Inc.

Norusis, M. (1988). SPSS-X Introductory Statistics Guide (for SPSS-X release 3). Chicago, IL: SPSS Inc.

Norusis, M. (1985). SPSS-X Advanced Statistics Guide. Chicago, IL: SPSS Inc.

Nunan, D. (1992). Research methods in language learning. New York: Cambridge University Press.

Oltman, P. K., Stricker, L. J.(1988). How native language and level of English proficiency affect the structure of the Test of English as a Foreign Language (TOEFL). East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED296592)

Ostler, S. (1980). A survey of academic needs for advanced ESL. TESOL Quarterly, XIV, 489-501.

Othuon, L. (1993). A study of the predictive validity of the Kenya certificate of primary education examination: Application of Hierarchical Linear Models. Unpublished

master's thesis, University of British Columbia, Vancouver, British Columbia, Canada.

Pack, A. C. A (1972). Comparison between TOEFL and Michigan Test scores and student success in (1) freshman and 2) completing a college program. TESL Reporter, 5, 1-7 & 9.

Parkerson, J. A., Lomax, R. G., Schiller, D. & Walberg, H. J. (1984). Exploring causal models of educational achievement. Journal of Educational Psychology, 76, 638-646.

Patkowski, M. S. (1991). Basic skills tests and academic success of ESL college students. TESOL Quarterly, 25, 735-738.

Pedhazur, E. (1982). Multiple regression in behavioral research. New York: CBS College Publishing.

Pedhazur, E & Schmelkin. (1991). Measurement, design, and analysis: Integrated approach. Hillsdale, NJ: LEA, Publishers.

Peirce, B. N. (1992). Demystifying the TOEFL Reading Test. TESOL Quarterly, 26, 665-691.

Perkins, K. (1988). Measuring ESL Readers' Ability to Apply Reasoning in Reading: A Validity Study of the TOEFL Reading Comprehension Subtest. Journal of Research in Reading, 11, 36-49.

Perkins, K. & Parish, C. (1988). What's wrong with reading comprehension tests? East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED297305)

Perry, W. S. (1989). The relationship of the Test of English as a Foreign language (TOEFL) and other critical variables to the academic performance of international graduate students. Dissertation Abstract International, 50, 422A.

Powers, D. E. (1985). A Survey of academic demands related to listening skills. Test of English as a Foreign Language Research Reports Number 20. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED304011)

Raimes, A. (1990). The TOEFL test of written English: cause for concern. TESOL Quarterly, 24, 227-243.

Reed, D. J. (1992). The relationship between criterion-based levels of oral proficiency and norm-referenced scores of general proficiency in English as a second language. System, 20, 329-345.

Ritsumeikan University. (1992). Ritsumeikan University (1992-1993). Kyoto, Japan: Ritsumeikan University.

Rivera, C. (Ed.) (1984). Language proficiency and academic achievement. Clevedon, England: Multilingual Matters Ltd.

Rosenthal, R. & Jacobson, L. (1968). Pygmalion in the classroom; teacher expectation and pupil's intellectual development. New York: Rinehart and Winston.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. TESOL Quarterly, 22, 69-90.

Sharon, A. T. (1972). English proficiency, verbal aptitude, and foreign student success in American graduate schools. Educational & Psychological Measurement, 32, 425-431.

Snow, C. E., Barnes W. S., Chandler J., Goodman I. F., & Hemphill L. (1991). Unfulfilled expectations: Home and social influences on literacy. Cambridge, MA: Harvard University Press.

Spolsky, B. (1990). The Prehistory of TOEFL. Language Testing, 7, 98-118.

Sposky, B. (1979). Some major tests. Arlington, Virginia: The Center for Applied Linguistics.

Sposky, B. (1978). Approaches to Language Testing. Arlington, Virginia: The Center for Applied Linguistics.

Sprinthall, N. & Sprinthall, R. (1994). Educational psychology: a developmental approach (6th ed.). New York: McGraw-Hill.

Swinton, S. S. & Powers, D. E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups. TOEFL Research Reports, 6. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED218921)

Tabachnick, B. G. & Fidell, L. S. (1989). Using multivariate statistics (2nd ed.). HarperCollinsPublishers, Inc.

Traynor, R. (1985). The TOEFL: An appraisal. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. EJ315491)

University of British Columbia. (1993). The University of British Columbia Calendar 1993-94. Vancouver, Canada: University of British Columbia.

Ushasree, S. (1990). Academic adjustment and scholastic achievement among the socially disadvantaged. Tirupati, India: Sri Venkateswara University.

Verhoeven, L. & Jong, J. (eds.) (1992). The construct of language proficiency: applications of psychological models to language assessment. Philadelphia: John Benjamins Publisher.

Vigotsky, L. (1986). Thought and language (Kozulin A., Trans.). Cambridge, MA: MIT Press.

Vinke, A. A. & Jochems, W. M. G. (1993). English proficiency and academic success in international postgraduate education. Higher Education, 26, 275-285.

Walberg, H. J. & Haertel, G. D. (eds.) (1990). The international encyclopedia of educational evaluation. New York: Pergamon Press.

Wan, T., Chapman. D. W., & Biggs, D. A. (1992). Academic achievement stress of international students attending U. S. universities. Research in Higher Education, 33, 607-623.

Willingham, W. M. (1990). Predicting college grades: An analysis of institutional trends over two decades. Princeton, NJ: Educational Testing Service.

Willingham, W. M. with Young, J. W., & Morris, M. M. (1985). Success in college: the role of personal qualities and academic ability. New York: College Entrance Board.

Wilson, K. M. (1982). A comparative analysis of TOEFL examinee characteristics, 1977-1979. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED237512)

Wilson, K. M. (1986). The relationship of GRE General Test scores to first-year grades for foreign graduate students: Report of a Cooperative Study. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED281862)

Wilson, K. M. (1987). Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED283831)

Wimberley, D. W., McCloud, D. & Flinn, W. (1992). Predicting Success of Indonesian Graduate Students in the United States. Comparative Education Review, 36, 487-508.

Wolf, F. (1986). Meta-analysis: Quantitative methods for research synthesis. Beverly Hills, CA: Sage Publications, Inc.

Yan, Z. (1994). Range restriction and indicator selection: a preliminary review of six prediction studies on relationship between TOEFL scores and GRE. Unpublished manuscript, University of British Columbia, Vancouver, British Columbia, Canada.

Yan, Z. (1994). A meta-analysis on studies of TOEFL scores' predictive validity on first year's GPA (1964-1994). Unpublished manuscript, University of British Columbia, Vancouver, British Columbia, Canada.

Young, J. W. (1993). Grade adjustment methods. Review of Education Research, 63, 151-165.

Yule, G., Hoffman, P. (1990). Predicting Success for International Teaching Assistant in a U. S. University. TESOL Quarterly, 24, 227-43.

Zeidner, M. (1987). A comparison of ethnic, sex, and age bias in the predictive validity of English language aptitude tests: some Israeli data. Language Testing, 4, 55-71.

Zirpoli, T. J., Halllahan, D. P., & Kneedler, R. D. (1988). The Indonesian project: correlates of student performance in a special education teachers training program. International Journal of Special Education, 3, 73-79.

Zhang S., Dunkel. P. & Zubovic, Y. (1992). Regression models in an ESL context: issues in construction and interpretation. TESOL Quarterly, 26, 191-196.

Appendix I The data file

	gpa	ened206a	ened379	educ395a	toefl	sec1
1	.	76.00	.	60.00	483.00	49.00
2	49.67	50.00	43.00	56.00	483.00	45.00
3	52.00	65.00	41.00	50.00	477.00	47.00
4	58.33	57.00	58.00	60.00	543.00	54.00
5	58.33	66.00	42.00	67.00	497.00	49.00
6	59.33	75.00	45.00	58.00	497.00	47.00
7	59.33	77.00	31.00	70.00	523.00	52.00
8	59.67	50.00	57.00	72.00	523.00	50.00
9	61.33	70.00	44.00	70.00	530.00	54.00
10	61.67	60.00	55.00	70.00	483.00	43.00
11	62.00	74.00	52.00	60.00	510.00	53.00
12	62.33	72.00	53.00	62.00	533.00	47.00
13	62.33	74.00	47.00	66.00	467.00	43.00
14	62.67	70.00	60.00	58.00	513.00	49.00
15	62.67	75.00	43.00	70.00	503.00	50.00
16	63.00	72.00	51.00	66.00	500.00	50.00
17	63.33	73.00	57.00	60.00	517.00	53.00
18	63.33	75.00	53.00	62.00	480.00	46.00
19	64.00	75.00	50.00	67.00	493.00	49.00
20	64.67	77.00	57.00	60.00	530.00	51.00
21	65.00	69.00	63.00	63.00	497.00	47.00
22	65.00	72.00	65.00	58.00	517.00	47.00
23	65.00	75.00	58.00	62.00	483.00	46.00
24	65.33	67.00	65.00	64.00	513.00	45.00
25	67.33	73.00	63.00	66.00	523.00	48.00
26	67.33	74.00	60.00	68.00	490.00	45.00
27	67.67	74.00	61.00	68.00	517.00	53.00
28	67.67	83.00	52.00	68.00	507.00	48.00
29	68.00	68.00	65.00	71.00	500.00	50.00
30	68.00	74.00	70.00	60.00	520.00	48.00
31	68.00	76.00	58.00	70.00	517.00	52.00
32	68.67	76.00	70.00	60.00	490.00	42.00
33	69.00	75.00	61.00	71.00	527.00	45.00
34	69.33	74.00	66.00	68.00	477.00	49.00
35	70.00	76.00	74.00	60.00	530.00	53.00
36	70.00	82.00	60.00	68.00	517.00	49.00
37	70.33	54.00	83.00	74.00	540.00	54.00
38	70.33	79.00	60.00	72.00	487.00	47.00
39	70.33	80.00	68.00	63.00	537.00	50.00
40	70.67	75.00	61.00	76.00	570.00	50.00
41	71.33	82.00	58.00	74.00	550.00	58.00
42	71.33	84.00	60.00	70.00	490.00	47.00
43	71.67	65.00	78.00	72.00	483.00	47.00
44	72.00	78.00	64.00	74.00	513.00	49.00
45	72.33	75.00	70.00	72.00	493.00	46.00
46	72.33	79.00	68.00	70.00	480.00	45.00
47	72.67	70.00	74.00	74.00	550.00	54.00
48	72.67	76.00	70.00	72.00	500.00	46.00
49	73.00	74.00	71.00	74.00	520.00	45.00
50	73.00	84.00	65.00	70.00	520.00	50.00
51	74.33	74.00	75.00	74.00	540.00	52.00
52	74.33	79.00	67.00	77.00	563.00	51.00
53	74.33	79.00	73.00	71.00	470.00	44.00
54	74.33	81.00	76.00	66.00	513.00	51.00
55	74.33	83.00	67.00	73.00	550.00	52.00
56	74.67	73.00	81.00	70.00	497.00	48.00
57	74.67	77.00	77.00	70.00	493.00	47.00

	gpa	ened206a	ened379	educ395a	toefl	sec1
58	74.67	77.00	77.00	70.00	557.00	54.00
59	74.67	78.00	76.00	70.00	543.00	49.00
60	74.67	82.00	70.00	72.00	473.00	46.00
61	74.67	84.00	67.00	73.00	553.00	55.00
62	75.00	77.00	78.00	70.00	527.00	51.00
63	75.33	79.00	75.00	72.00	470.00	45.00
64	75.33	82.00	73.00	71.00	477.00	44.00
65	75.33	84.00	74.00	68.00	507.00	47.00
66	75.67	77.00	78.00	72.00	540.00	56.00
67	75.67	79.00	71.00	77.00	553.00	55.00
68	75.67	86.00	77.00	64.00	503.00	52.00
69	75.67	89.00	71.00	67.00	503.00	48.00
70	76.00	79.00	76.00	73.00	510.00	48.00
71	76.00	82.00	72.00	74.00	550.00	56.00
72	76.33	76.00	81.00	72.00	500.00	41.00
73	76.67	75.00	83.00	72.00	497.00	47.00
74	76.67	78.00	80.00	72.00	540.00	56.00
75	77.00	76.00	77.00	78.00	487.00	39.00
76	77.00	82.00	77.00	72.00	507.00	49.00
77	77.67	74.00	85.00	74.00	540.00	51.00
78	77.67	81.00	78.00	74.00	503.00	50.00
79	78.67	79.00	86.00	71.00	543.00	52.00
80	79.00	83.00	77.00	77.00	520.00	48.00
81	79.00	90.00	77.00	70.00	540.00	50.00
82	79.33	82.00	78.00	78.00	533.00	52.00
83	79.67	88.00	78.00	73.00	520.00	47.00
84	80.00	81.00	81.00	78.00	543.00	50.00
85	80.00	81.00	83.00	76.00	513.00	52.00
86	80.00	85.00	78.00	77.00	533.00	51.00
87	82.33	75.00	94.00	78.00	480.00	48.00
88	82.33	86.00	80.00	81.00	530.00	50.00
89	82.33	87.00	83.00	77.00	580.00	61.00
90	82.67	83.00	85.00	80.00	573.00	61.00
91	82.67	85.00	83.00	80.00	573.00	56.00
92	82.67	88.00	80.00	80.00	513.00	52.00
93	82.67	89.00	81.00	78.00	497.00	47.00
94	83.00	82.00	80.00	87.00	550.00	59.00
95	83.33	86.00	84.00	80.00	530.00	54.00
96	83.33	90.00	83.00	77.00	560.00	53.00
97	87.00	90.00	87.00	84.00	543.00	56.00

	sec2	sec3	write	speak	gender	toeflpre
1	48.00	48.00	2.00	.90	1.00	483.00
2	50.00	50.00	1.00	.90	1.00	483.00
3	48.00	48.00	2.00	.90	1.00	477.00
4	55.00	54.00	3.00	1.90	1.00	543.00
5	47.00	53.00	1.00	1.00	1.00	497.00
6	51.00	51.00	3.00	1.90	2.00	497.00
7	53.00	52.00	1.00	1.90	1.00	507.00
8	54.00	53.00	2.00	.90	1.00	523.00
9	57.00	48.00	3.00	1.90	1.00	530.00
10	48.00	54.00	3.00	.	1.00	483.00
11	52.00	48.00	2.00	1.90	1.00	510.00
12	58.00	55.00	4.00	1.90	2.00	533.00
13	50.00	47.00	2.00	.90	1.00	467.00
14	52.00	53.00	3.00	.90	2.00	513.00
15	52.00	49.00	2.00	2.00	1.00	503.00
16	52.00	48.00	3.00	1.90	1.00	500.00
17	48.00	54.00	3.00	2.00	1.00	517.00
18	49.00	49.00	2.00	2.00	2.00	480.00
19	52.00	47.00	3.00	1.00	1.00	460.00
20	54.00	54.00	2.00	2.00	1.00	530.00
21	53.00	49.00	2.00	1.00	2.00	497.00
22	55.00	53.00	2.00	1.90	1.00	517.00
23	51.00	48.00	3.00	.00	2.00	483.00
24	55.00	54.00	2.00	1.00	1.00	513.00
25	58.00	51.00	3.00	1.00	1.00	523.00
26	51.00	51.00	2.00	.90	1.00	490.00
27	49.00	53.00	1.00	1.00	2.00	517.00
28	52.00	53.00	3.00	.90	1.00	507.00
29	51.00	49.00	2.00	.00	2.00	500.00
30	54.00	54.00	2.00	.00	1.00	520.00
31	50.00	53.00	4.00	2.00	1.00	517.00
32	55.00	50.00	4.00	1.00	1.00	490.00
33	61.00	52.00	2.00	.90	1.00	487.00
34	47.00	47.00	2.00	.90	1.00	477.00
35	52.00	54.00	3.00	2.00	2.00	530.00
36	53.00	53.00	2.00	1.00	1.00	517.00
37	55.00	53.00	4.00	1.90	2.00	540.00
38	50.00	49.00	2.00	.90	1.00	487.00
39	56.00	55.00	3.00	1.90	1.00	523.00
40	68.00	53.00	3.00	1.00	2.00	513.00
41	56.00	51.00	3.00	2.00	2.00	550.00
42	52.00	48.00	2.00	2.00	2.00	490.00
43	46.00	52.00	2.00	1.00	1.00	483.00
44	54.00	51.00	3.00	1.90	2.00	483.00
45	54.00	48.00	2.00	.90	2.00	493.00
46	51.00	48.00	2.00	1.90	1.00	480.00
47	61.00	50.00	4.00	3.00	2.00	547.00
48	53.00	51.00	3.00	1.90	1.00	487.00
49	58.00	53.00	4.00	1.90	1.00	520.00
50	56.00	50.00	5.00	2.00	2.00	520.00
51	54.00	56.00	2.00	1.90	1.00	520.00
52	58.00	60.00	2.00	.	1.00	547.00
53	52.00	45.00	2.00	1.00	2.00	470.00
54	54.00	49.00	3.00	1.90	2.00	493.00
55	54.00	59.00	3.00	1.00	1.00	550.00
56	51.00	50.00	3.00	1.00	2.00	497.00
57	51.00	50.00	3.00	1.00	2.00	493.00

	sec2	sec3	write	speak	gender	toeflpre
58	57.00	56.00	2.00	1.90	1.00	557.00
59	60.00	54.00	2.00	1.90	1.00	517.00
60	52.00	44.00	4.00	2.00	2.00	473.00
61	55.00	56.00	3.00	1.90	2.00	553.00
62	54.00	53.00	2.00	1.90	2.00	527.00
63	48.00	48.00	4.00	1.00	2.00	470.00
64	52.00	47.00	3.00	1.90	1.00	477.00
65	55.00	50.00	3.00	.00	2.00	507.00
66	53.00	53.00	3.00	2.00	2.00	540.00
67	56.00	55.00	2.00	1.00	2.00	553.00
68	51.00	48.00	2.00	1.90	1.00	503.00
69	52.00	51.00	2.00	2.00	2.00	503.00
70	53.00	52.00	3.00	.00	2.00	510.00
71	56.00	53.00	4.00	1.90	2.00	550.00
72	53.00	56.00	2.00	1.00	2.00	500.00
73	52.00	50.00	3.00	1.00	2.00	497.00
74	58.00	48.00	3.00	.90	2.00	520.00
75	53.00	54.00	4.00	2.90	1.00	487.00
76	52.00	51.00	2.00	1.00	2.00	507.00
77	61.00	50.00	2.00	1.90	2.00	540.00
78	49.00	52.00	3.00	1.90	2.00	503.00
79	55.00	56.00	4.00	2.00	2.00	543.00
80	56.00	52.00	3.00	.90	2.00	520.00
81	57.00	55.00	4.00	1.90	1.00	520.00
82	56.00	52.00	3.00	2.00	2.00	493.00
83	56.00	53.00	2.00	1.00	2.00	513.00
84	59.00	54.00	2.00	1.00	1.00	543.00
85	52.00	50.00	3.00	1.00	2.00	513.00
86	58.00	51.00	3.00	1.90	1.00	533.00
87	49.00	47.00	3.00	.90	2.00	480.00
88	56.00	53.00	2.00	.90	1.00	530.00
89	57.00	56.00	3.00	2.00	2.00	577.00
90	58.00	53.00	4.00	2.00	2.00	570.00
91	58.00	58.00	3.00	1.00	2.00	573.00
92	50.00	52.00	3.00	2.00	2.00	513.00
93	52.00	50.00	2.00	.00	2.00	497.00
94	54.00	52.00	3.00	2.00	2.00	550.00
95	55.00	50.00	4.00	2.00	2.00	530.00
96	61.00	54.00	4.00	2.00	2.00	550.00
97	53.00	54.00	3.00	3.00	2.00	543.00

Appendix II The list of standardized residuals and leverage values

	gpa	zre_tot	lev_tot	zre_sec	lev_sec
1	78.67	.01185	.03357	.00005	.05956
2	62.33	-.26058	.04696	-.27775	.05135
3	75.33	.37326	.08908	.45587	.09910
4	75.67	-.16919	.02028	-.02946	.03603
5	67.67	-.18236	.03353	-.16574	.04249
6	74.67	.76297	.04023	.53343	.08277
7	82.67	1.65650	.06361	1.63646	.06392
8	52.00	-2.02214	.03501	-1.87327	.05101
9	76.67	-.05763	.03317	.00203	.11375
10	75.00	.14203	.03923	.09495	.04372
11	74.33	.32776	.06275	.39459	.11503
12	65.00	-.43412	.02645	-.55406	.04046
13	82.67	.26183	.06595	.41617	.09953
14	65.00	-1.09487	.02780	-1.16161	.03486
15	83.33	.91557	.03011	.98745	.04238
16	79.67	.93115	.02610	.74468	.05455
17	71.33	-.98196	.02740	-.86176	.05509
18	62.67	-.62034	.03393	-.56548	.04415
19	59.33	-2.20797	.02870	-2.22204	.03879
20	61.33	-1.43484	.02255	-1.36927	.09307
21	72.67	.73991	.02744	.68385	.03040
22	49.67	-2.21252	.05536	-2.22396	.05557
23	72.33	1.19734	.05094	1.15341	.05441
24	65.33	-.36852	.02076	-.52355	.04633
25	80.00	1.47875	.02394	1.40383	.04503
26	83.33	.53302	.04774	.38734	.06668
27	76.00	.18139	.07434	.18329	.07680
28	76.67	.49493	.02110	.47693	.02220
29	80.00	.81801	.01580	.92246	.02695
30	70.33	-1.27192	.03182	-1.20212	.03535
31	69.00	.02855	.03199	-.27853	.12271
32	67.67	-.67309	.07477	-.52036	.10099
33	79.00	.97624	.06004	.92182	.06622
34	77.67	.39796	.04508	.15958	.13103
35	58.33	-2.07501	.03101	-1.98387	.03685
36	75.67	1.43117	.02962	1.56725	.06390
37	74.33	-.03998	.01664	-.04228	.02403
38	76.33	.65906	.02644	.38595	.16597
39	68.00	-.70106	.06366	-.60169	.07461
40	82.67	.47460	.08217	.47333	.09221
41	74.33	.72534	.05533	.61600	.08467
42	75.67	.55774	.05063	.48508	.05638
43	87.00	1.62708	.07591	1.71437	.09016
44	74.67	.22967	.02394	.23312	.02561
45	71.67	1.01556	.02852	1.17619	.07769
46	74.67	.58444	.05824	.57255	.05847
47	82.67	1.28205	.01993	1.39658	.04128
48	62.00	-.81926	.02711	-.67128	.07029
49	73.00	.28247	.05445	.08020	.08797
50	67.33	-.43488	.03378	-.54908	.05677
51	74.33	.74850	.03733	.76494	.04961
52	79.00	.56638	.01995	.44869	.03077
53	69.33	.71885	.03501	.94238	.07997
54	63.33	-1.09931	.07792	-1.11695	.08167
55	59.33	-1.15162	.07445	-1.11406	.07686
56	82.33	.37496	.07149	.51442	.08633
57	64.00	-.57943	.03394	-.47375	.06678

	gpa	zre_tot	lev_tot	zre_sec	lev_sec
58	63.00	-.78872	.02744	-.68818	.04791
59	64.67	-.64837	.03318	-.64558	.03547
60	72.33	.11155	.03066	-.01559	.05153
61	83.00	.86294	.02740	1.05386	.06134
62	71.33	.03825	.06359	-.03909	.07055
63	65.00	-1.21379	.07931	-1.18508	.08110
64	68.67	-.05833	.08838	-.20536	.10605
65	72.67	-.98388	.07231	-1.11125	.12319
66	68.00	-.07843	.08129	-.08717	.08378
67	72.00	-.40896	.01664	-.47273	.01968
68	68.00	-.46571	.05530	-.24876	.10158
69	70.33	.74946	.02684	.82353	.03234
70	67.33	.23680	.02513	.21706	.02765
71	77.67	.61466	.02305	.71146	.05036
72	80.00	1.56824	.04706	1.42202	.06560
73	77.00	1.37816	.12488	1.11990	.24115
74	75.33	.11421	.07353	.04687	.08050
75	77.00	.67523	.02460	.67218	.02519
76	70.00	-.93777	.01693	-.85039	.03661
77	62.67	-1.92721	.01889	-1.90177	.03147
78	73.00	-.84607	.08662	-.86892	.08959
79	59.67	-1.39631	.02875	-1.37712	.02915
80	76.00	-.50334	.03817	-.41273	.04435
81	79.33	.49989	.01754	.44542	.02009
82	58.33	-1.01632	.04976	-.85125	.08982
83	74.67	.27193	.09043	.24740	.11580
84	74.67	.17866	.02110	.21727	.02329
85	75.67	-.12219	.05862	-.10667	.05977
86	70.00	.31844	.02230	.33727	.02368
87	82.33	1.60356	.03890	1.70789	.05079
88	82.33	2.09875	.03482	2.04925	.03743
89	70.33	-.10089	.02631	-.14970	.03334
90	74.67	-.49733	.02887	-.45606	.04251
91	63.33	-.94856	.02250	-.68273	.09999
92	62.33	-2.44774	.02963	-2.65863	.08243
93	75.33	1.45488	.05424	1.39603	.06382