# TOEIC® RESEARCH REPORT

Number 3

November 1998

# Scores on the TOEIC® (Test of English for International Communication) Test as a Function of Training Time and Type

Robert F. Boldt and Steven J. Ross

THE CHAUNCEY GROUP INTERNATIONAL™

**TOEIC Research Report Number 3**

# Scores on the TOEIC® (Test of English for International Communication) Test as a Function of Training Time and Type

Robert F. Boldt & Steven J. Ross

The Chauncey Group International
Princeton, New Jersey

# Scores on the TOEIC® (Test of English for International Communication) Test as a Function of Training Time and Type


## Abstract

The TOEIC® (Test of English for International Communication) test has been popular among companies since its introduction in 1979 for assessing English in professional and business contexts. In 1997 more than 1.5 million people took the TOEIC test worldwide. The TOEIC test is used to evaluate the English language proficiency of candidates for training and assignment as organizations become increasingly involved with people from other nations.

Training is an important context for the use of the TOEIC test. A decision to train employees involves not only an assessment of their current levels of proficiency, but also predictions of the levels of proficiency they might reach given additional language training. These predictions require some understanding of the effectiveness of the type, length, purpose, and intensity of the language training program.

The purpose of this study was to better understand the factors influencing the effectiveness of English language training. This was accomplished by combining and analyzing existing sets of TOEIC test scores and related data to obtain results that were more powerful than could be obtained with separate analyses.

Data sets were obtained from 23 Japanese companies and training institutions. The institutions supplied pre- and post-training Reading Comprehension and Listening Comprehension TOEIC scores, as well as information on training time, course materials, instructor qualifications, training objective, and class size. A total of 4,247 trainee records were used in the data analysis.

The total sample was split into two portions composed of 1/3 and 2/3 of the total sample. Preliminary analyses of the smaller sample determined the linear formulas appropriate for predicting post-training performance, as well as the variables that should be included in the formulas.

The larger sample was used for simulating the flow of trainees through English language training programs and into a pool of potential assignees to jobs requiring a certain level of English proficiency. The simulations showed that the materials used and the instructor qualifications had the largest effects on the number of trainees reaching the required level of English proficiency and on the average post-training TOEIC score of these trainees.

# TABLE OF CONTENTS

## LIST OF TABLES AND FIGURES

# Scores on the TOEIC® (Test of English for International Communication) Test as a Function of Training Time and Type

## Introduction

The TOEIC® (Test of English for International Communication) test is used by more than 2,000 organizations worldwide for assessing English in professional and business contexts. The TOEIC test has been popular among companies and language schools since its introduction in 1979. This popularity has continued to grow as organizations increasingly pursue a global market strategy that includes the need to communicate in English. In 1997 over 1.5 million people took the TOEIC test in more than 20 countries.

English language skills are a factor in many human resource decisions, such as employee recruitment and selection, placement, and career advancement. Even after a candidate is hired, language training is often necessary before a job assignment can be finalized. Clearly, an assessment of English language proficiency is needed, and the TOEIC test has been used extensively to fulfill that need.

The decision to train employees involves not only an assessment of existing levels of proficiency, but also a prediction of whether training can bring employees up to an adequate level of proficiency and an estimate of how much training will be needed to do so. These predictions require some understanding of the effectiveness of the type, length, and intensity of training.

In 1993, the TOEIC Research Advisory Panel addressed the need of many companies to know the amount of instruction required to raise the English language proficiency of their employees to specific TOEIC score bands. A program of controlled research that isolated the effects of all variables affecting training outcomes would have required daunting amounts of time and money. Fortunately, existing data sets containing TOEIC test data and related training data were available and were merged for use in this study. Analyses based on this combined data, initially gathered in many independent data sets, were used to obtain results that are potentially more powerful than those obtained through separate analyses. These data sets provided baseline information on post-training performance as a function of length, intensity, and types of training. A preliminary survey of the availability of the data needed for this study indicated that the number of data sets and the means to collect them were available only in Japan. Hence this study dealt entirely with the effects of training on TOEIC scores in Japanese organizations.

## Course Variables

### Materials

English-as-a-Foreign-Language materials vary considerably depending on the focus of the course. For the most part, modern commercial materials are "communicative" in that they encourage learners to use their language skills to communicate meaning. This contrasts with the "structural" materials characteristic of the 1960's and 70's, which tended feature linguistic systems analyzed in terms of their constituent parts. The focus at that time was on building proficiency from the bottom up through the use of sequential, structured materials in a "building block" manner.

The materials sampled in the current study were mostly communicative. They differed mainly in their degree of communicative orientation and focused on featuring English language in different contexts of usage. Learners were presumed to have a pre-existing schematic reference point allowing them to recognize how linguistic forms fit into a specific context of use. Some, for instance, featured language use in the context of on-the-job interaction. Other materials were not contextualized in any given job-specific format and presumably relied on learners' world knowledge in the most general sense. The

purpose of this genre of materials was to help students associate linguistic forms with their own communicative intentions about issues anticipated to be of popular interest.

In this study, the classification of materials was based on two major criteria. The first was the client organization's identification of the type of materials used in the particular courses featured in the study. The second was, in some cases, a corroborating review of the actual course material. This system led to a general classification of the teaching material. It should be noted, however, that some client organizations used a multi-faceted approach to materials selection. The emphasis in the classification exercise was to identify the predominant type of materials used in the featured courses.

**Features of Training Materials**

The following paragraphs outline the major features of the training materials used. The materials are described in terms of primary language syllabus organization principles (e.g., structural, functional, task-based), and in terms of their potential overlap with the TOEIC test content domain.

***Business Simulation (BusSim).*** This popular approach to language pedagogy for employees is based on the process of simulation. Materials are devised to match a context in which the language learners are most likely to use the target language. Business simulations often present role play scenarios in which learners attempt to use the target language in a plausible "real-life" context. An example is an overseas client who has sent a faxed message complaining that previously ordered spare parts have not yet arrived. The learners simulate a telephone call in which they act out both the roles of the overseas client as well as the company employee responsible for explaining the shipping problem.

Business simulations often follow a deductive learning scheme. Sample contexts are presented in the form of dialogues, which are then practiced. Interactional rules and patterns are presented and then practiced intensively. Role plays are repeated until the learners are reasonably confident that they can apply the formulas in comparable contexts in the work place. Such simulations are often problem-solving exercises crafted to feature common communication problems encountered by international employees.

The interface of business simulations to the linguistic content of the TOEIC test is indirect. Attention to form in business simulations may be less overt than in a structural syllabus. The major connection between business simulation materials and the TOEIC test is in the domain of functional language use. As the test is written to sample the universe of language forms occurring in business communication, simulations potentially overlap with test content specifications in terms of the contexts of use, and sometimes in the structural characteristics of language used in business.

***General English (GenEng).*** The classification of materials as "General English" is the most variable in this study. General English texts differ considerably in quality and domain sampling. Most texts of this genre are published by multinational publishing concerns and marketed internationally. Most of these texts are designed to be used "off the shelf" by minimally trained native-speakers of English. A few require in-house training seminars to orient new teachers to orthodox use of the materials.

The large majority of English-as-a-Foreign-Language teaching materials feature a heterogeneous structural "backbone" organized by language functions. For instance, a lesson segment might feature language functions such as 'promising' or 'making an appointment'. Different ways to promise or make an appointment might be contrasted. In this regard, general English materials may differ from the business simulations in that the former sample the most common contexts of use, while the latter would seek to sample language functions in the context of business interactions. General English texts also usually feature some form of language structure analysis. For the communicative functions of 'promising' or 'making an appointment', auxiliary verbs or modals might be introduced as well. These, for instance, would feature semantic differences between 'will' (a definite commitment) and 'could' (indicating possibility). General English materials thus provide a wider sampling of

contexts of use, but a more narrow focus and language structure.

In texts for general English, the four skills of reading, writing, listening, and speaking are usually featured. The typical implementation of general English texts in company programs tends to weigh listening and speaking more heavily than reading and writing.

Compared to business simulation materials, general English materials may better match the domain of language sampled by the TOEIC test. This coverage is mainly due to the range of sampling of English structure. While some of the contexts of interaction featured in the general English texts have little overlap with the content domain of international English (e.g., a 'process' lesson in which A instructs B in the process of making an omelet), a wider sampling of language structures, and relatively more frequent use of written texts increases its potential overlap with the two skills of reading/structure and listening assessed on the TOEIC test.

***Current Events/News (News).*** This relatively unstructured approach to language teaching material features current events for reading and discussion. The content coverage of these materials, which are for the most part English language newspapers and magazines such as Newsweek or Time, differs from both business simulations and general English. Newspapers and current events magazines are not edited for non-native readers, and thus may contain idiomatic and specialized usage not common to international English, or to business domains of language use. No direct interface to language functions is provided in these news articles.

The usual approach to using current events is for learners to read and discuss the news articles. The purpose of these materials is often language maintenance rather than the systematic learning of new structures or functions. The use of current events articles can nonetheless lead to incidental learning of new vocabulary and grammatical forms.

There is considerable interface of current events and news materials to the content of the TOEIC test for the reading/structure section. However, because the input from current events materials is primarily visual, learners' experience with listening tasks like those featured on the TOEIC test is probably limited. The content of discussions about current events may be conducive to improved fluency in oral communication, but whether this is reflected on listening gain is a question for further research.

***Video materials (Video).*** This increasingly popular mode of language instruction involves the use of video-taped vignettes of 'real-life' interaction. These provide contextualized input to language forms and functions through the audio-visual channel. Depending on the design of the video materials—whether they are devised to be for English-as-a-Foreign-Language pedagogy, or are samples of authentic language for non-pedagogical purposes—the salience of form and function can differ. Materials designed to teach forms and functions make these features of language optimally salient through the scripted interactions of the actors. Authentic materials, which are more common, make no such attempt.

Video materials differ from the current events and news materials mainly in that current events and news are more frequently presented in written format. Video materials, whatever their linguistic organization, require aural and visual processing. While current events materials may provide a relatively rich sampling of language structure and vocabulary in printed form, the video materials provide contextual clues about the meaning of language forms and functions through the visual medium. It is likely, therefore, that there would be a differential interface between these two types of materials with the content and format of the TOEIC test.

The effective use of video materials may enhance learners' listening comprehension skill. With little printed input, their reading skill may well be less engaged by this type of materials. As noted earlier, this differential focus is often intentional; for example, many Japanese learners have greater reading and grammar skill development than aural comprehension at the beginning of instruction. The choice to enhance what is least developed is often a pragmatic one, commonly driven by organizational needs.

The interface of video materials to the TOEIC test appears to be the opposite of that for the current events and newspaper articles. If the input is mainly audio-visual, it is likely that aural comprehension would be most affected. The listening comprehension portion of the TOEIC test could therefore be expected to be more strongly related to gains due to the use of video materials.

**Instructor Variables**

***Bachelor's degree (BA).*** Many programs and language schools employ young university graduates who are native English speakers. The most common degree or qualification observed in the sample was a bachelor's degree. The major field was not specified, as it is not considered important as a hiring criterion in the majority of company programs sampled.

***In-house programs (PRG).*** Some language programs provide training courses for their own language teachers. These training courses typically provide neophyte teachers who are native speakers of English with strategies for using specified course materials or teaching methods, often devised on an in-house basis. The chief purpose of in-house programs is to ensure some uniformity in the use of the materials sold as part of the language training package. The training regimen rarely includes an outline of theories of cognitive development, language acquisition, or assessment, and provides mainly "how to" overviews of teaching techniques.

***Certificate (CRT).*** Completion of language teaching certification courses is not uncommon among expatriate language teachers. These courses usually provide an orientation to practical teaching techniques for native speakers of English. While the range of curriculum is certainly much broader than in the typical in-house training scheme, the certification courses focus mostly on practical classroom management techniques, correction or feedback methods, syllabus organization principles and assessment techniques.

***Master's degree (MA).*** In Japan, instructors holding master's degrees in any field are still relatively rare in language schools and company language training programs. The possible ambiguity in encoding a bachelor's degree in any field as a single label also applies to the coding of master's degrees. Some bachelor's degrees, for instance, may be directly relevant to English language teaching (e.g., modern languages or linguistics), while some master's degrees may be in areas not directly applicable to language teaching (e.g., finance or geography). The specific details of graduate training in a particular field were, unfortunately, not available from company or language school program administrators, thus making for potential inferential problems in observing differences in effects attributable to educational qualifications.

**Course Objectives**

One reason companies provide language training programs is to promote the English language development of their employees. In some programs, employees with needed technical expertise are given language training regardless of their language aptitude or prior experience in language learning. This phenomenon is related to the globalization of many corporations and the increased need for English language skills to facilitate communication with colleagues and clients around the world, either in person (as in overseas postings) or by telephone, fax, e-mail, etc.

In contrast, language schools and college or university programs are less influenced by specific outcome expectations. The focus of these programs is the development of overall language proficiency. The coding system used in this study relates to the stated goal of the program as specified by the program administrators.

***General Education (GenEd).*** The most common objective for all of the types of programs sampled in the survey was to develop general English language proficiency. This goal corresponds approximately to the use of materials covering all four skill areas.

***Staff Development (StaffDev).*** Company programs differ from language schools or colleges and universities in that they aim to increase the potential for individual employees to work in a wide variety of job assignments. One manifestation of

this strategy is to provide extensive language instruction to all employees, regardless of whether specific plans exist to post them overseas or to give them job assignments requiring proficiency in English. Company programs attempt to create a pool of employees from which the company may subsequently draw persons with English language skills in order to meet future organizational needs.

***New Employees (NewEmp).*** Company programs differ from language schools in another aspect: new employees may be assessed with the TOEIC test in order to evaluate the need for further language training. Once these needs are identified, new employees may be given intensive language training in-house. This strategy indicates a company policy of employee evaluation prior to more permanent job assignments in the home country or overseas. In contrast with the staff development objective, the new employee training objective is seen as a pre-assignment evaluation of the employee's readiness for different kinds of jobs required in the organization.

**Class Size**

Language class size varied considerably in the programs sampled in this study. Class size has a clear potential impact on the rate of gain in language development, since the amount and focus of input provided to individuals can be expected to vary with different teacher-to-student ratios. In order to assess the impact of class size on language gain (as measured by the TOEIC test), class size was broken down as follows:

***Large***. Classes with more than 20 students.
***Mid***. Classes with between 11 and 20 students.
***Small***. Classes with 10 or fewer students.

Class sizes corresponded to program types. College and university programs typically had large classes (about 35 students per class). Company and language school programs were primarily in the middle to small class range, depending on the program objectives. Extensive courses, in which instruction was given over several months, were mostly mid-sized. Intensive courses tended to have small classes.

**Data Sets**

Data sets were obtained from 23 Japanese companies and training institutions. Each data set included information on from 30 to 1,030 trainees. The language proficiency test data supplied were pre- and post-training Reading Comprehension and Listening Comprehension TOEIC scores, except in one case where only TOEIC total scores were available. Training time comprised total number of weeks and total number of hours. In most of the analyses the training time figures used were those intended for the course and did not include actual attendance time. Other training information included: materials used in the training (general English materials, newspapers, video materials, or business simulation); qualifications of the instructors (general Bachelor's degree, any Master's degree, certificate in teaching, or training in specific teaching materials); objective of the training (general education, staff development, or new employee); and class size (greater than 20, from 11 to 20, and 10 or fewer). For each type of information, only one category was reported for each subject. For example, even though both video materials and newspapers may have been used in the same course, only the predominant material was identified (e.g., video or newspapers, but not both).

Because the participating organizations had not collected all relevant information for the present study, and because some time had passed since the data collection had occurred, there was considerable variation across individual data sets in the percent of cases for which information was missing. The main analysis was conducted using the 4,247 cases for which the Listening Comprehension score, the Reading Comprehension score, the number of weeks of training, and the total hours of training were specified, and for which the class size was greater than 10 (although only one organization had class sizes larger than 20). This sample will be referred to as the "major sample." The amount of missing data was an important factor in determining this sample; the major sample was constructed in order to provide symmetric cell sizes in later analyses.

The major analyses in the study included two steps: (1) estimation of constants (such as regression weights or treatment effect sizes), and (2) evaluation of prediction results obtained using these constants. These two steps should use data sets that contain different but comparable cases. Therefore, the major sample was divided into two sub-samples. The first sub-sample (the "estimation" sample) was used for estimating the constants; the second sub-sample (the "cross" sample) was used for evaluating prediction results. The two sub-samples were formed by removing every third case from the major sample, resulting in an estimation sample with 1,416 cases and a cross sample with 2,831 cases.

## Analyses

The purpose of the present study was to develop formulas to predict post-training performance given pre-training test and course data. Both linear and non-linear formulas were studied (see Appendix A for details about the non-linear formulas examined). It was found that the linear prediction system was easier to understand, easier to use, was less prone to validity shrinkage in a new sample, and predicted as well as the non-linear formulas. For these reasons, the use of non-linear prediction was not considered further.

Linear regression was used to predict post-training test scores. The purpose of the regression analysis was to develop a weighted composite of predictor scores and a set of adjustments to the composite to account for course variables. The weights and adjustments in the composite were chosen so as to maximize correlation of the composite with post-training scores. This analysis used the 1,416-case estimation sample.

***Imputation***. In some cases, no record or memory of course conditions was available. In other cases, information was missing. Therefore, it was necessary to impute conditions to replace the missing data. Imputations were made for course materials, instructor requirements, and class size. The purpose of the training (e.g., GenEd) was known for all cases so no imputation was necessary.

"General English" was used as the default course material when there was no actual record of the material used. Other possible imputations were News, Video, and BusSim. The use of both video and business simulation requires substantial preparation; it was felt that use of these materials would be remembered by our informants even though some time had passed since the delivery of the courses. Therefore, in the absence of specific knowledge of the use of video or a business simulation, it seemed wisest to assume that general English materials were used.

"Bachelor's degree (BA)" was used when information about the instructor qualification was missing. The complete range of qualifications of the instructors included a general Bachelor's degree, any Master's degree, certificate in teaching, or training in specific teaching materials. Of these, only BA and training in specific teaching materials (PRG) occurred in the major sample. Since the training in specific teaching materials would require substantial preparation and would have been remembered by our informants, it seemed wisest to impute BA as the instructor requirement when the requirement was not known.

"Mid" was used when class size information was missing. For only two companies were class sizes known to be less than 10 or greater than 20; the rest of the class sizes were known to be intermediate or were unknown. It seemed wisest to impute an intermediate size to classes of unknown size.

With the imputations described above, the sample took a 3x2x3x2 complete factorial design. BusSim, GenEng and Video were the levels of the first factor, which was the materials factor; BA and PRG were the levels of the second factor, which was the instructor requirement; GenEd, StaffDev, and NewEmp were the levels of the third factor, which was the course purpose; and Mid and Large made up the fourth factor, which was class size. The treatment cells did not have equal or proportional replication, so the significance test for the treatment effects associated with course variables followed a procedure described by Kempthorne (1952) for treating the general p-way classification without interaction.

***Errors of prediction***. A realistic prediction system should take the following two factors into

account: First, prediction formulas are more accurate for estimation samples than for cross samples. Second, the actual post-training scores are "distributed around," rather than "equal to," the value calculated using the prediction formula; hence, the prediction should be probabilistic. Both of these factors were taken into account by using the cross sample to develop a distribution of errors of prediction.

## Results

When using a prediction formula, one must develop empirical constants that are appropriate to the specific class of data for which predictions are being made. The well-known regression weight is an example. In traditional linear prediction studies, constants are estimated using an estimation sample, while holding out a similar sample for evaluation of predictive accuracy (Cohen & Cohen, 1983). In this study, a 33% spaced sample was separated from the total data set in order to evaluate the prediction results.

***Use of test scores and training time in composites.*** The first decision made was whether both Listening and Reading test scores were needed, and whether all of the number of weeks, the total course hours, and "intensity" were needed. "Intensity" was the dividend of total course hours divided by the number of weeks. Various combinations of these variables, plus dummy variables indicating the company from which the data had been acquired, were evaluated in the trial prediction functions.

The results for predicting the post-training Listening score are given in Table 1. The following abbreviations are used in all tables throughout the paper: Listening pre-training score ($L_1$), Reading pre-training score ($R_1$), Total pre-training score ($T_1$), number of weeks of instruction (Wks), total number of hours of instruction (Hrs), and intensity (Int). Dummy variables indicating the company are included in all composites. The likelihood ratio criterion statistical procedure (Dobson, 1983; Stuart, 1991) was used to compare results across predictor sets.

Table 1 shows that the multiple correlations ($R$) of the predictor sets with post-training Listening scores were very similar. However, the large sample size ($N$ = 1,416) provided enough power to distinguish some differences. If a significant difference occurs when a variable is dropped from the prediction set, it indicates that the variable is needed to more accurately predict the outcome (e.g., post-training Listening scores). If the difference is not significant, the variable does not contribute to the predication of the outcome and should not be included in the prediction set.

The chi-square ($\chi^2$) results presented in Table 1 indicate that when the pre-training Reading score was dropped from the prediction set it resulted in a significant loss of prediction (comparisons 1 vs. 2 and 3 vs. 4). Therefore, the pre-training Reading score was included in the composite predicting the post-training Listening score.

**Table 1.** Multiple and squared multiple correlations between composites based on predictor sets and the post-training Listening score.

| | Predictor Set | $R^2$ | $R$ | Comparison | $\chi^2$ |
|---|---|---|---|---|---|
| 1. | $L_1$,$R_1$,Wks,Hrs,Int | .6366 | .7979 | | |
| 2. | $L_1$,Wks,Hrs,Int | .6259 | .7911 | 1 vs. 2 | 41.362* |
| 3. | $L_1$,$R_1$,Hrs | .6363 | .7977 | 1 vs. 3 | 1.402 |
| 4. | $L_1$,Hrs | .6256 | .7909 | 3 vs. 4 | 41.127* |
| 5. | $L_1$,$R_1$ | .6359 | .7974 | 3 vs. 5 | 1.663 |
| | Prediction Composite: Reading score, Listening score | | | | |

* $p < .001$

Testing result 5 against result 3, and result 3 against result 1, provided a surprise. These comparisons tested whether training time contributed significantly to the accuracy with which post-training Listening scores were predicted. The chi-square for this comparison was not significant, indicating that the total number of weeks, total number of hours, and training intensity were not needed in a composite for predicting the post-training Listening scores. Only the pre-training Reading score (and pre-training Listening score) was necessary in the prediction composite.

Results for predicting the post-training Reading score are given in Table 2, which, as with Table 1, reveals that the correlations were very similar, but that some differences were detectable. Table 2 shows that the pre-training Listening score made a significant contribution to the accuracy of prediction of the post-training Reading score and so was included in a composite predicting the post-training Reading score. The total number of weeks of training and the training intensity did not make a significant contribution to the accuracy of predicting the post-training Reading score and were not included in the prediction formula.

Testing result 5 against result 3 in Table 2 led to a different outcome from that in Table 1. Again, this test assessed whether any of the variables describing training time made a significant contribution to the accuracy of prediction of post-training Reading scores. In this case, the total number of hours of training did indeed make a contribution to the accuracy with which the post-training Reading score was predicted. Accordingly, the composite for predicting post-training Reading scores included the Listening scores and the total number of hours of training.

Results for predicting the Total score are presented in Table 3. One issue when forming the composite for predicting post-training Total scores was whether the pre-training Listening and Reading scores should be used separately, or whether the pre-training Total score was sufficient. Table 3 shows that when the Listening and Reading scores were replaced with the Total score, no loss in predictive accuracy resulted (comparisons 1 vs. 2 and 3 vs. 4). Hence, the pre-training Total score was used in the prediction composite. Neither weeks of training nor intensity of training were needed in the prediction composite because dropping them from the prediction set yielded a non-significant chi-square.

Finally, it seemed logical to conclude that, since the total hours of training contributed to the accuracy of prediction of post-training Reading scores and since Reading makes up a substantial portion of the Total score, total hours of training should be a predictor of the post-training Total score. When the hours of instruction were dropped from the predictor set, it resulted in a significant

**Table 2.** Multiple and squared multiple correlations between composites based on predictor sets and the post-training Reading score

| | Predictor Set | $R^2$ | $R$ | Comparison | $\chi^2$ |
|---|---|---|---|---|---|
| 1. | $L_1$,$R_1$,Wks,Hrs,Int | .7045 | .8394 | | |
| 2. | $R_1$,Wks,Hrs,Int | .6958 | .8341 | 1 vs. 2 | 41.305* |
| 3. | $L_1$,$R_1$,Hrs | .7044 | .8393 | 1 vs. 3 | .446 |
| 4. | $R_1$,Hrs | .6956 | .8340 | 3 vs. 4 | 41.632* |
| 5. | $L_1$,$R_1$ | .7028 | .8383 | 3 vs. 5 | 8.081* |

Composite: Listening score, Reading score, total hours of training

* $p < .001$

**Table 3.** Multiple and squared multiple correlations between composites based on predictor sets and the post-training Total score.

| | Predictor Set | $R^2$ | $R$ | Comparisons | $\chi^2$ |
|---|---|---|---|---|---|
| 1. | $L_1$,$R_1$,Wks,Hrs,Int | .7281 | .8533 | | |
| 2. | $T_1$,Wks,Hrs,Int. | .7280 | .8532 | 1 vs. 2 | .854 |
| 3. | $L_1$,$R_1$,Hrs | .7281 | .8533 | 1 vs. 3 | .158 |
| 4. | $T_1$,Hrs | .7279 | .8532 | 2 vs. 4 | .155 |
| | | | | 3 vs. 4 | .851 |
| 5. | $T_1$ | .7268 | .8525 | 4 vs. 5 | 6.024* |
| Composite: Total score, total hours of training | | | | | |

* $p < .02$

loss of predictive accuracy. Accordingly, both the pre-training Total score and the total hours of training were used in the composite for predicting the post-training Total score.

***Estimating and testing course descriptor effects.*** Dummy variables were used to index the course conditions used in training (see Appendix B for a discussion of dummy variables). A basic combination of conditions was used as a baseline effect to which the other effects were added. Course conditions making up the basic combination included:

- The use of general English materials
- BA required for the instructor
- General education as the course objective
- A class size of 10 to 20 students

The value of the dummy variable for the basic combination was set at one for every case, with other dummy variables set at one only if the condition for the case differed from the basic combination. For example, to calculate the combined effects of course conditions for a person who was trained using video for staff development purposes in a class of moderate size, the weight for video and the weight for staff development were added to the weight for the basic combination (i.e., the dummy variables for the basic combination, video, and class development were one and all other dummy variables were zero). No adjustment for class size was needed because a moderate sized class was a condition of the basic combination. Had the trainee been a new employee or had the class size been large, further adjustments using the weights for new employees

**Table 4.** Significance tests of course effects in the prediction of post-training Listening, Reading, and Total scores

| | Listening Scores[a] | | Reading Scores[b] | | Total Scores[d] | |
|---|---|---|---|---|---|---|
| **Effects** | *d.f.* | *F* | *d.f.* | *F* | *d.f.* | *F* |
| BusSim & Video | 2,1407 | 20.482** | 2,1406 | 5.555** | 2,1407 | 9.235** |
| PRG | 1,1407 | 16.891** | 1,1406 | .365[c] | 1,1407 | 4.842* |
| StaffDev & NewEmp | 2,1407 | 31.970** | 2,1406 | 15.668** | 2,1407 | 30.919** |
| Large Class | 1,1407 | 4.652* | 1,1406 | 6.196* | 1,1407 | 2.088[e] |

* $p < .05$, ** $p < .01$

[a] Predictors were $L_1$, $R_1$ and dummy variables.

[b] Predictors were $L_1$, $R_1$, Hrs, and dummy variables.

[c] After dropping PRG $R^2 = .7044$ and $R = .8393$

[d] Predictors were $T_1$, Hrs, and dummy variables.

[e] After dropping Large Class $R^2 = .7275$ and $R = .8530$

and large class size would have been necessary.

Table 4 contains the results of testing course effects in the prediction of post-training test scores. Examination of Table 4 revealed that the course materials and purposes consistently had significant effects on the accuracy of prediction of post-training test performance. However, the PRG result indicated an effect of in-house instructor training on post-training Listening scores, but not on Reading scores. Also, while the Large Class effect was significant for both Reading and Listening, it was not significant for the Total score.

***Small class results.*** Data for small classes were available from only two companies, one of which was unable to furnish Listening and Reading scores. Thus the Small Class results for Total that follow use data from two companies, but results for Listening and Reading are given for the one company for which both Listening and Reading data were supplied.

In these analyses the values of the previously determined treatment effects were used, and the analyses estimated and tested treatment effects for those treatments not appearing in previous analyses, (i.e., News, MA, CRT, and Small Class). Only the Small Class effect could be tested for Listening and Reading. The results were *t*-tests for Listening ($t_{(277)}$ = -.6707, ns) and Reading ($t_{(277)}$ = -23.5851, $p < .01$), respectively.

As has been mentioned, one company supplied only Total scores. In addition to using a small class size, this company also used newspapers as the principal teaching material, and required an MA for some instructors and a certificate for others. Hence, several effects could be estimated and tested for Total score prediction. These results are presented in Table 5, in which all tests indicated statistical significance.

***Magnitude of treatment effects.*** The significance tests above indicate the presence or absence of treatment effects but do not indicate the magnitude of those effects. Table 6 presents the treatment effects. The entries pertaining to test scores or total hours (i.e., the entries for Listening, Reading, Total score, and total hours) are small relative to other entries because these variables are covariates and are to be multiplied by their respective observations to estimate post-training test scores. All other entries are treatment effects and are on the scale of the post-training test score. The entry for the basic combination in Table 6 applies to every score; all entries below the basic combination are additions to that value.

One notable feature of Table 6 is that when new employees were being trained a substantially larger post-training score was forecast than when the purpose of the course was general education or staff development. Note also the unexpected non-significant or negative values for the Small Class effects, and the positive value for the Large Class effect when predicting post-training Listening scores.

Table 6 shows that the effects of the various learning conditions were not the same for predicting post-training Listening and Reading scores. For example, the effect of Video was approximately 17 points more for Listening than for Reading. PRG increased the predicted post-training Listening score by approximately 32 points but had no significant effect on Reading scores.

**Table 5.** Significance tests of small class course effects in prediction of post-training Total scores[a]

| Effects | *d.f.* | *F* |
|---|---|---|
| News | 1, 750 | 24.702** |
| MA & CRT | 2, 750 | 8.831* |
| Small | 1, 750 | 27.830** |

* $p < .05$, ** $p < .01$

[a] The small class results are based on 754 individuals from 2 companies.

Finally, note the small value of the weights for total hours as compared to those for the Listening, Reading, and Total scores. This occurs partly because the test scores have the greater predictive value, and partly because the scales of the variables differ. That is, the standard deviation of Listening and Reading scores was 91 and 87, respectively, while the standard deviation of total hours was 170. The weight for the variable with the larger standard deviation must be smaller if the products of the weights and their covariates are to be on the same scale (i.e., the scale of the post-training test score). The standard deviation of the Total score was 147, nearer to the standard deviation of the total hours, which emphasizes the greater role of the pre-training test score than of total hours in predicting the post-training test score.

***Distribution of the errors of prediction.*** Because the actual post-training scores were "distributed around," rather than "equal to," the value calculated using the prediction formula, it was necessary to investigate the distribution of the errors of prediction. This investigation revealed that, relative to variation in the predicted and actual scores, the variances of the error of estimate were homogeneous for the Listening, Reading, and Total score data (see Appendix C). Examination of the distribution of errors of prediction also suggested that errors of prediction might be pooled when making probabilistic predictions of post-training scores. This would be a considerable simplification in the prediction system, and the possibility of a further simplification of the assumption of normality errors of prediction was

**Table 6.** Estimated treatment effects for post-training Listening, Reading, and Total scores

| Effect | Listening ($L_i$) | Reading ($R_i$) | Total |
|---|---|---|---|
| Listening Comprehension | .525 | .163 | NS |
| Reading Comprehension | .173 | .589 | NS |
| Total score | * | * | .732 |
| Total hours | NS | .112 | .076 |
| **Basic combination** | **87.009** | **52.326** | **133.493** |
| ***Course Materials*** | | | |
| BusSim | 18.458 | 16.937 | 38.162 |
| Video | 31.756 | 14.94 | 53.207 |
| ***News*** | ** | ** | 44.207 |
| ***Instructor Background*** | | | |
| PRG | 32.242 | NS | 35.287 |
| CRT | ** | ** | 36.800 |
| MA | ** | ** | 5.351 |
| ***Course Objective*** | | | |
| StaffDev | 15.517 | 11.856 | 29.880 |
| NewEmployee | 59.451 | 39.598 | 103.078 |
| ***Class Size*** | | | |
| Small class | NS | -65.583 | -14.641 |
| Large class | 10.262 | -23.913 | NS |

NS indicates a non-significant effect after accounting for the basic combination or for use of the total score.

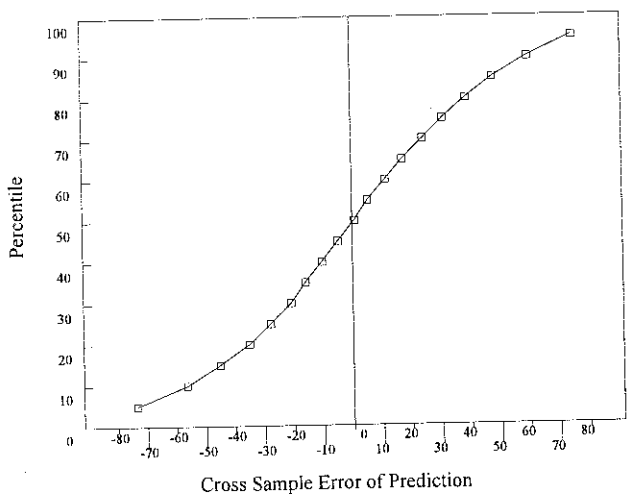
* Listening and Reading scores were used separately instead of the Total score.

** News, CRT, and MA were not used except in the sample where only Total scores were available for small classes. On lines where ** appears, the Total effect was estimated using cases trained in small classes (see text).

explored. To do this, the skewness and kurtosis of the errors of prediction were calculated (see Appendix D for further details). However, all but one of the skewness coefficients were significant, violating the assumption of normality. Hence, the normal distribution was not adopted for use with the errors of prediction.

Figure 1 presents a graph of the cumulative percentile distribution of errors of prediction for the post-training Listening score. The Y-axis shows the percent of test-takers actually scoring below a given predicted score. The X-axis shows the difference in actual TOEIC scores (in either a positive or negative direction) from a predicted score. The errors of prediction used to produce Figure 1 were computed in the cross sample using parameters from the estimation sample, and pooled across all estimated score levels from 200 to 425. This score range represents the largest observed range in Listening scores for the sample used in this study. Figure 1 does not apply to scores below 200 or above 425 on the Listening score scale.

**Figure 1:** Prediction Error Percentiles – Listening Comprehension




If we assume, for example, a predicted Listening score of 300, Figure 1 shows that approximately 50% of those with estimated post-training Listening scores of 300 would have errors of prediction of zero or less (i.e., 50% would have post-training Listening scores of 300 or less). Similarly, approximately 20% of those with an estimated Listening score of 300 can be expected to attain post-training Listening scores at least 35 points below 300 (i.e., below 265). If we now assign a "passing score" of 335, Figure 1 shows that, of those individuals with a predicted score of 300, 80% would actually obtain a score below 335. In other words, only 20% of test-takers would "pass" the test.

Figure 1 can be used to help interpret the quantities in Table 6. For example, suppose that the 300-point estimate for Listening was obtained using the basic combination of course characteristics, which includes using general English instruction materials. In Table 6 we find that using Video materials adds approximately 32 points to the estimate over that afforded if only the basic combination was used. The added 32 points would increase the predicted score from 300 to 332. Still assuming a "passing score" of 335, Figure 1 shows that just over 50% would actually score below 335. In effect, the pass rate has increased from 20% to just under 50%. This assumes, of course, that the course was reoriented around the Video materials, and that the introduction was not simply a pro forma change. Since the effect of using instructors prepared through an in-house program (PRG) was also about 32 points, the introduction of such an instructor instead of Video could be expected to have a similar effect quantitatively. The reader should keep in mind, however, that the 30% gain was only for those whose estimated score was 300; the quantitative effect for a class whose pre-training estimated scores varied even though those scores averaged 300.

Figures 2 and 3 have interpretations similar to that of Figure 1 but for Reading and Total scores, respectively. Examination of Figures 5 and 6 reveals that approximately 20% of test-takers will score at least 30 points below their predicted Reading score and 50 points below their predicted Total score. Examination of Table 5 reveals that no change of course material or teacher requirement will bring estimated scores of 300 near the hypothesized passing score of 335, since the changes for BusSim and Video were approximately 17 and 15 points, respectively, for post-training Reading scores. However, Figure 3 suggests

that the introduction of video would bring the pass rate up to the 50% level for people predicted at the 20% rate. That is, examination of Figure 3 reveals that the 20th percentile lies at about -50 on the abscissa in Figure 3, and in Table 6 we see that the introduction of Video would raise the estimated score a little over 50 points.

**Figure 2:** Prediction Error Percentiles – Reading Comprehension

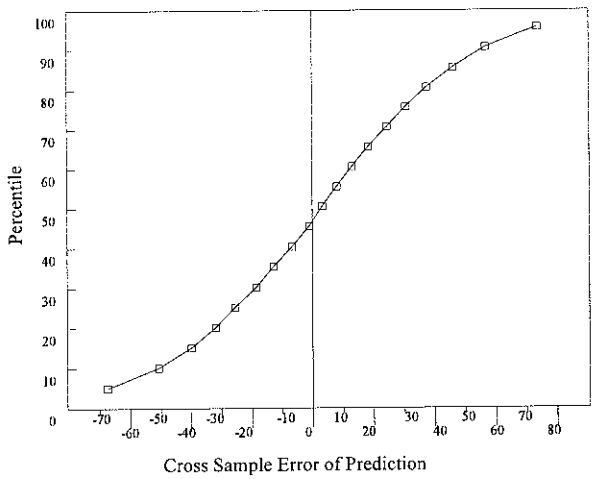
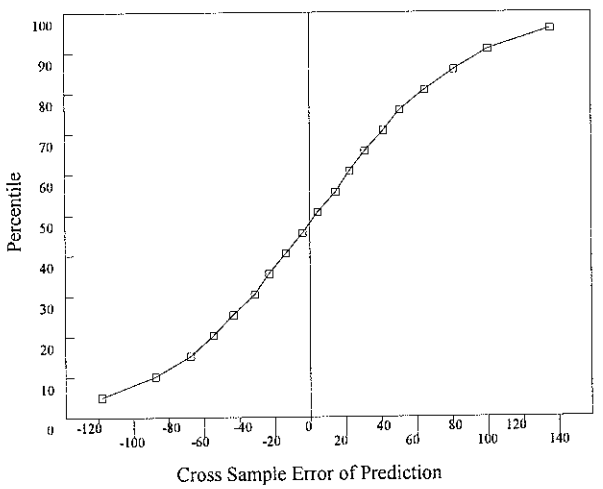



**Figure 3:** Prediction Error Percentiles – Total Score




**Simulation of Training Outcomes**

The parameters in Table 6 and the cumulative distributions of errors were used to simulate the aggregated effects of pre-training test scores, cut score requirements for training entry, and course conditions on post-training TOEIC test scores. The simulations reported here are given both for their own interest and as examples of the types of results that could be generated using the data and programs developed in this project. The following steps were required:

(1) Identify the group from which trainees were to be selected,
(2) Set the policy for selecting trainees and identify a minimum required post-training score, and
(3) Identify the course conditions to be used.

The simulation identified those individuals selected for training, calculated the simulated post-training scores for those selected, and ran descriptive statistics on trainees and assignees.

For this study, all trainees in the cross sample were selected. It should be noted, though, that the simulation could be operated using as input any sub-set of the cross sample, or perhaps cases from some other similar sample so long as the TOEIC scores, total training time, and course conditions were known and were compatible with the conditions used here.

The simulation required a minimum pre-training Total TOEIC score for selection to training, and a minimum post-training TOEIC score. In order to add a degree of realism to the simulation, we sought minimum scores currently in use in business. Two simulations were carried out using different sets of pre- and post-training minimum scores.

Several sets of course conditions were used for the simulations. The parameters in Table 6 provided estimates of the contribution of each course condition to the prediction of training outcomes. For example, Table 6 shows that the Total test score and total training hours had weights of .732 and .076, respectively. The additive constant for the basic combination was

133.493. Thus, for the basic combination, Formula 1 below yields the estimated post-training Total score as a function of the pre-training Total score and the total hours.

**Formula 1.**

Predicted post-training Total score = .732 [pre-training Total score] + .076 (total hours) + 133.493

Because the basic combination served as a baseline from which to calculate the contribution of all other course conditions, Formula 2 was used to estimate the post-training Total score when business simulation materials replaced general English materials.

**Formula 2.**

Predicted post-training Total score = (value of Formula 1) + 38.162

That is, the formula for predicting post-training success when business simulation materials replaced general English materials was equal to Formula 1, which applies to the basic combination, plus the parameter in Table 6 for business simulation (38.162). This is so because all parameters other than the basic combination were calculated as additions to the basic combination.

In a third example, video materials were used in class and participation in a special training program was required of the instructor for an intermediate sized class of new employees. In this case, the formula for predicting the post-training Total score was

**Formula 3.**

Predicted Post-training Total score = (value of Formula 1) + 53.207 + 35.287 + 103.078

The numbers in Formula 3 come from the Table 6 values for Video (53.207), PRG (35.287), and NewEmployees (103.078). Nothing was added for the intermediate class size condition because it was a feature of the basic combination. Other combinations of the conditions listed in Table 6 could be used; their effect would be to change the value to be added to Formula 1.

Post-training prediction formulas, such as those given above, comprised only one part of the simulated post-training TOEIC scores. The errors of prediction (i.e., differences between actual and predicted scores) were used to account for the fact that actual post-training scores were distributed around, rather than equal to, the value calculated using the prediction formula. Because a number of errors of prediction were observed, a great number of actual post-training scores could arise from any particular predicted score. To obtain a simulated post-training score, one error value was randomly selected and added to the predicted post-training score derived from the prediction formula.

Pre-training TOEIC scores were used to simulate the number of individuals selected for training. Simulated post-training scores were used to compute the number and average post-training test scores of those meeting a minimum post-training cut-off score. The use of course conditions to determine the prediction formulas allowed simulating the effects of these conditions on the estimated and hence the simulated post-training TOEIC scores.

Simulation programs are subject to the effects of random selection of errors, and thus the results can differ over several simulations that use the same conditions. This instability was reduced in the present study by averaging the results over 500 simulations.

Tables 7 and 8 present simulation results for several combinations of course conditions. The tables are divided into sections pertaining to the purposes of training (i.e., general development, staff development, and new employees). All of the course conditions were manipulated in these tables except for the TOEIC scores and the number of hours of training.

Results for News, CRT, MA and class size are not given in Tables 7 and 8. The Table 6 results for these course conditions were estimated on only a few cases and pre-training TOEIC scores for Listening and Reading were not available for News, CRT, and MA. Also, the results for class size were unexpected. Substantial negative treatment effects for small classes, such as those found for Reading in Table 6, were counterintuitive.

**Table 7.** Number qualified, Number passing[a], pass rates[ab], and predicted post-training TOEIC Total Score Means[ab] and Standard Deviations[ab] (SD) for various combinations of course conditions and a Total score of 650 for selection for training.

| Course Conditions | No. Qual.[c] | No. Pass | Pass Rate%[d] | Total | | Listening | | Reading | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean[e] | SD[e] | Mean[e] | SD[e] | Mean[e] | SD[e] |
| ***Results for General Development*** | | | | | | | | | |
| Basic[f] | 372 | 88.5 | 24 | 810.1 | 57.2 | 417.0 | 39.7 | 393.1 | 36.5 |
| BusSim | 372 | 144.0 | 39 | 817.2 | 59.4 | 420.6 | 40.4 | 396.6 | 37.6 |
| Video | 372 | 169.1 | 45 | 821.1 | 60.7 | 430.1 | 40.9 | 391.0 | 38.4 |
| PRG | 372 | 139.3 | 37 | 816.6 | 58.9 | 435.6 | 40.2 | 380.9 | 37.5 |
| BusSim and PRG | 372 | 204.3 | 55 | 827.8 | 63.6 | 441.6 | 42.1 | 386.2 | 39.5 |
| Video and PRG | 372 | 230.4 | 62 | 832.4 | 64.9 | 451.5 | 42.8 | 380.9 | 40.2 |
| ***Results for Staff Development*** | | | | | | | | | |
| Basic | 372 | 130.8 | 35 | 815.3 | 58.8 | 420.8 | 40.2 | 394.4 | 37.5 |
| BusSim | 372 | 194.8 | 52 | 825.5 | 62.4 | 426.3 | 41.7 | 399.2 | 39.0 |
| Video | 372 | 221.1 | 59 | 831.0 | 64.8 | 436.5 | 42.5 | 394.6 | 40.1 |
| PRG | 372 | 189.8 | 51 | 824.5 | 62.6 | 441.3 | 41.6 | 383.2 | 39.1 |
| BusSim and PRG | 372 | 255.1 | 69 | 839.1 | 67.9 | 448.8 | 43.9 | 390.2 | 41.4 |
| Video and PRG | 372 | 278.1 | 75 | 846.3 | 70.2 | 460.2 | 45.0 | 386.1 | 42.4 |
| ***Results for New Employees*** | | | | | | | | | |
| Basic | 372 | 254.7 | 68 | 838.9 | 67.5 | 439.8 | 43.6 | 399.1 | 41.0 |
| BusSim | 372 | 307.7 | 83 | 858.6 | 73.9 | 450.2 | 46.4 | 408.4 | 43.7 |
| Video | 372 | 323.4 | 87 | 868.1 | 76.2 | 462.6 | 47.5 | 405.5 | 44.8 |
| PRG | 372 | 304.3 | 82 | 857.1 | 73.2 | 464.8 | 46.0 | 392.3 | 45.5 |
| BusSim and PRG | 372 | 339.8 | 91 | 882.2 | 79.3 | 478.2 | 48.7 | 404.0 | 46.1 |
| Video and PRG | 372 | 348.6 | 94 | 893.6 | 81.1 | 491.7 | 49.5 | 401.8 | 47.1 |

[a]Calculations assume that all those qualified were trained.
[b]Each set of means and standard deviations was based on 500 simulation runs, i.e., 18 sets of 500 runs in all.
[c]A pre-training Total score of 650 was required for selection.
[d]The base for the pass rate was the number qualified for training; the minimum post-training TOEIC score for passing was 750.
[e]Means and standard deviations (SD) were for TOEIC Total scores.
[f]Basic combination: teacher with BA, mid-sized class, general English texts, class held for general development

Given the already considerable size and complication of Tables 7 and 8, and the fact that the size effects were based on small samples, we chose not to include the size effects in the simulation.

The simulation results presented in Table 7 are based on a subsample of 372 cases, each of which had a minimum pre-training TOEIC score of 650. The post-training "pass score" was set at 750. Table 7 shows, for example, that, of the 372 trainees who received training under the basic combination for the purpose of general development, 88.5, or 24%, equaled or exceeded the post-training standard of 750, and that the post-training TOEIC total scores of those exceeding the standard had a mean of 810.1 and a standard deviation of 57.2.

Table 7 also shows differences in post-training test scores and pass rates attributable to various course

conditions. For example, using video materials resulted in a mean post-training test score of 821.1 and a 45% pass rate, double that produced by the basic combination (i.e., introducing video into the classroom resulted in double the pass rate obtained in a mid-sized class held for reasons of general development, taught by a teacher with a BA and using general English texts). This implies that enhancing the basic combination with video would yield the same number of trainees with post-training scores over 750 as would training twice as many employees using the basic combination alone.

Table 7 also allows comparisons across course objectives. For example, using video in courses for general development and staff development resulted in increases in the pass rate of 21% and 24%, respectively, over that due to the basic combination. Using video with new employees resulted in only a 19% improvement in the pass rate. Determining whether the smaller improvement in pass rate achieved by using video with new employees would be worthwhile involves evaluating the added cost of using video.

***Effects of Variation in Training Time.*** The total hours of training observed in this study ranged from 12 to 441 (mean = 211.8 hours, median = 147 hours). Thus the component of the predicted post-training TOEIC Total score that was due to training hours was at most 33.5 points (441 hours x .0732, where .0732 is the effect of training hours on total scores from Table 6). The modal number of hours training was 368 (1684 cases with 58% of the cases being trained fewer hours).

Note also in Table 6 that introducing a video into the training adds 53.207 points to the predicted total score. Comparing this figure with the 33.5 points from the paragraph above reveals that the gain from doubling the largest number of hours of training would not match the addition to the predicted post-training TOEIC Total score that would be expected from introducing the video. This holds true regardless of the purpose for which training was given. The reader can verify that similar calculations for introducing BusSim, News, a training program for the instructor, or CRT would lead to a similar result.

***Effects of Variation in Minimum Scores.*** Table 7 was calculated using 650 as the minimum score for selection to training and 750 as the minimum post-training "pass score". Other scores might be used. For example, Anderson Worldwide, a leading international consulting firm, identifies five job levels from "Analyst" to "Partner" with minimum TOEIC scores ranging from 620 for Analyst to 800 for Partner (The Reporter, No. 23, 1997). The score of 750 that was used in Table 7 was the minimum score for the second highest level at Anderson Worldwide (Associate Partner). Also, the TOEIC representative in Japan some time ago developed a five-level proficiency scale with the levels-score combinations as follows: A-860, B-730, C-470, and D-220, and a lowest level labeled E. The scale was assembled using informally collected Japanese experience and anecdotal data that related TOEIC scores to levels of proficiency. The consensus at the time was that a score of 600 was considered the minimum level of proficiency in English necessary for conducting international business. It was apparent from the Anderson Worldwide and the TOEIC Proficiency Scale levels that the minimum standards of 650 and 750 are fairly high. The C level of the TOEIC Proficiency Scale is described as representing "Sufficient knowledge for daily activities and conducting business within certain limits." To show the effect of variation in minimum scores, Table 8 was developed in the same way as Table 7 except that the minimum score for selection to training was taken as 220, the bottom of the D level, and the minimum "pass score" was taken as 470, the bottom of the C level.

In contrast with Table 7, the numbers selected for training, the numbers qualifying for assignment, and the pass rates were all larger, reflecting the easing of the test-based standards. All of the pass rates were above 50%, making it impossible for variations in course conditions to affect training yields as much as was possible when the high standards of Table 7 were used. Note that when the training was for general education the pass rate was 52%, rising to 66% with the use of video, which was far from the doubling found in Table 7. However, in Table 8 the introduction of Video to the training for general education resulted in an additional 381 trainees qualified for assignment, whereas the corresponding number in

**Table 8.** Number qualified, Number passing[a], pass rates[ab], and predicted post-training TOEIC Total, Listening, and Reading Score Means[ab] and Standard Deviations[ab] (SD) for various combinations of course conditions

| Course Conditions | No. Qual.[c] | No. Pass | Pass Rate %[d] | Total | | Listening | | Reading | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean[e] | SD[e] | Mean[e] | SD[e] | Mean[e] | SD[e] |
| ***Results for General Development*** | | | | | | | | | |
| Basic[f] | 2633 | 1365.2 | 52 | 597.5 | 95.0 | 312.4 | 53.8 | 285.0 | 54.9 |
| BusSim | 2633 | 1636.6 | 62 | 611.3 | 102.6 | 320.4 | 57.1 | 290.9 | 58.6 |
| Video | 2633 | 1746.7 | 66 | 617.0 | 105.7 | 331.2 | 58.5 | 285.8 | 60.2 |
| PRG | 2633 | 1615.7 | 61 | 610.2 | 102.1 | 335.4 | 56.8 | 274.8 | 58.4 |
| BusSim and PRG | 2633 | 1895.5 | 72 | 624.8 | 110.0 | 343.8 | 60.3 | 281.0 | 62.1 |
| Video and PRG | 2633 | 2003.7 | 77 | 631.3 | 113.1 | 355.1 | 61.8 | 276.2 | 63.8 |
| ***Results for Staff Development*** | | | | | | | | | |
| Basic | 2633 | 1576.4 | 60 | 608.2 | 100.9 | 319.9 | 56.4 | 288.3 | 57.7 |
| BusSim | 2633 | 1856.0 | 70 | 622.6 | 108.7 | 328.2 | 59.7 | 294.4 | 61.5 |
| Video | 2633 | 1965.3 | 75 | 628.7 | 111.8 | 339.2 | 61.3 | 289.6 | 63.1 |
| PRG | 2633 | 1834.8 | 70 | 621.6 | 108.2 | 343.3 | 59.5 | 278.4 | 61.3 |
| BusSim and PRG | 2633 | 2107.2 | 80 | 637.7 | 116.1 | 352.4 | 63.1 | 285.4 | 65.0 |
| Video and PRG | 2633 | 2203.9 | 84 | 645.1 | 118.9 | 364.0 | 64.5 | 281.1 | 66.5 |
| ***Results for New Employees*** | | | | | | | | | |
| Basic | 2633 | 2105.2 | 80 | 637.6 | 116.1 | 343.2 | 63.1 | 294.4 | 64.7 |
| BusSim | 2633 | 2331.8 | 89 | 657.8 | 123.3 | 354.1 | 66.6 | 303.7 | 68.0 |
| Video | 2633 | 2401.0 | 91 | 667.1 | 125.8 | 366.7 | 67.8 | 300.4 | 69.3 |
| PRG | 2633 | 2317.3 | 88 | 656.1 | 122.9 | 368.9 | 66.3 | 287.2 | 68.0 |
| BusSim and PRG | 2633 | 2475.9 | 94 | 681.1 | 128.8 | 382.3 | 69.2 | 298.8 | 70.7 |
| Video and PRG | 2633 | 2517.5 | 96 | 692.5 | 130.8 | 396.1 | 70.0 | 296.5 | 71.8 |

[a]Calculations assume that all those qualified were trained.
[b]Each set of means and standard deviations was based on 500 simulation runs, i.e, 18 sets of 500 runs in all.
[c]A pre-training Total score of 220 was required for selection.
[d]The base for the pass rate was the number qualified for training; the minimum post-training TOEIC score for passing was 470.
[e]Means and standard deviations (SD) were for TOEIC Total scores.
[f]Basic combination: teacher with BA, mid-sized class, general English texts, class held for general development

Table 7 was 80. Other comparisons also reveal that the smaller gains in pass rates in Table 8 (as compared to those in Table 7) were offset by the large number of cases trained in Table 8.

Examination of the mean TOEIC Total scores revealed that those means varied directly with the number of trainees who qualified for assignment (No. Qual.). This was a direct result of the fact that both figures depend on the mean estimated post-training score, which was in turn determined by the value of the treatment effect appearing in Table 6. However, the differences in the overall size of the data entries between Tables 7 and 8 were quite large as compared to the variation within the tables. Thus the level of the minimum standard can be a major factor in determining the yield of training in terms of quantity and quality of those available for assignment.

***Effects on TOEIC Listening and TOEIC Reading.*** The minimum standards used in the simulations were TOEIC Total standards. However, selection and training have effects on Listening and

Reading scores as well, and simulation results for these scores are also presented in the tables. As with the Total score, the simulated post-training Listening and Reading scores were computed using predicted post-training scores to which randomly selected errors of prediction were added. The formulas for predicted post-training Listening and Reading scores were similar to those for the Total, with the regression weights and treatment effects taken from the appropriate columns in Table 6.

For example, the formula for the predicted post-training score for Listening as a function of pre-training Listening and Reading scores when using the basic combination was

**Formula 4a:**

Predicted post-training Listening score = .525(Listen) + .173(Read) + 87.009

and the corresponding formula for the post-training Reading score using pre-training Listening and Reading scores and Total hours was

**Formula 4b:**

Predicted post-training Reading score = .162(Listen) + .589(Read) + .076(Total hours) + 52.326

where the constants in (4a) and (4b) come from the first and second data columns under Listening and Reading, respectively, of Table 6.

When Video was added to the basic combination the formula for predicting post-training TOEIC Listening was

**Formula 5a:**

Predicted post-training Listening score = (Value of Formula 4a) + 31.756

and the formula for predicting post-training TOEIC Reading was

**Formula 5b:**

Predicted post-training Reading score = (Value of formula 4b) + 14.945

where 31.756 comes from the second data column of Table 6 and 14.945 comes from the third data column of Table 6. Thus the process of constructing predicted post-training performance on TOEIC Listening and Reading was analogous to that for TOEIC Total except for the columns in Table 6 from which the constants were taken.

As with the TOEIC Total, errors of prediction were computed for Listening and Reading by subtracting predicted post-training scores from actual post-training scores. Thus, each case contributed a vector of three errors of prediction, one for Listen, one for Read, and one for Total. These vectors were assembled into a file of errors with three columns. Whenever a simulation was in process, the error portion of a vector of simulated actual post-training scores would consist of a randomly selected row from the file of errors. The simulated actual post-training scores for a case were created by calculating the predicted post-training scores using formulas (4a and b) and (5a and b) above, and adding the vector selected from the file of errors. The resulting vector of three scores was then adjusted as outlined below.

One additional adjustment to the simulated Listening and Reading scores was necessary. The need for this adjustment arose because the computational procedure described above did not retain the property that simulated Total scores should be equal to the sum of simulated Listening and simulated Reading scores. This important constraint on the scores was imposed by a formula that produced simulated post-training Listening and Reading scores that summed to the Total score.

Means and standard deviations of the post-training TOEIC scores are given in the right-hand section of Tables 7 and 8. The means for Reading were substantially less than the means for Listening. This was not a feature produced by the simulation but a feature of the data available in which the average of the Listening scores exceeded the average of the Reading scores for both the pre- and post-training scores. For the total sample, the average pre-training Listening and Reading scores were 240.4 and 199.6, respectively; for the post-training Listening and Reading scores, the averages were 283.8 and 232.1 respectively.

Another salient feature of Tables 7 and 8 was that all of the standard deviations in Table 8 were larger than the corresponding standard deviations in Table 7. For example, the standard deviation of TOEIC Listening for new employees was 63.1 for Table 8 and 43.6 for Table 7. This difference occurred because the higher selection and assignment standards imposed in Table 8 restricted the range of scores selected for training and for post-training assignment. Note, however, that every standard deviation for Listening in Table 7 exceeded the corresponding standard deviation for Reading in that same table. The opposite is true in Table 8; the standard deviations for Reading exceed the corresponding standard deviations for Listening.

As might be expected, the relationship of Listening and Reading standard deviations in Table 8 was more similar to that in the cross sample than that in Table 7, in that the standard deviation for Reading (83.7) exceeded that of the standard deviation for Listening (79.1). The effect of the simulation then was to restrict the standard deviation for Reading more than for Listening. This could result from the higher rate of screening, which affected the Reading scores more than the Listening scores because more of the Reading scores were low. The data in Tables 7 and 8 suggest that changing the selection and assignment standards did indeed bring the Listening and Reading means closer. Taking the differences between Listening and Reading means reveals that for every combination of course conditions the larger value comes from Table 8, which is what one would expect if the selection excluded more of those scoring low on Reading.

Finally, Tables 7 and 8 reflect the greater contribution of Video and PRG to the Listening than to the Reading treatment effects. For example, in Table 7, when moving from the second to the third then to the fourth data line, the means for Listening can be seen to increase (420.6, 435.1, 441.6) while those of Reading decrease (396.6,391.9, 330.2). Parallel trends occur in all three sections of both Tables 7 and 8 and are a consequence of similar trends in the treatment effects shown in Table 6.

The impact of video use on listening gain may be related to the fact that video-based teaching materials provide rich and authentic input via the aural channel. Utterances are contextualized in a manner that facilitates the pragmatic mapping of meanings onto situational correlates of utterances. The influence of PRG on Listening is more general. PRG encodes training in the use of particular instructional materials. To the extent that materials are systematically used by instructors in a program, less variance can be expected in the use of the course materials. Course impact on listening development, in other words, can be expected to be more homogeneous in instructional programs that include in-service training modules for instructors.

## Discussion

This study examined the amount and kind of training needed to produce significant changes in TOEIC scores. Answering this question rigorously would require access to pre- and post-training TOEIC scores where the training varied in length, where the course conditions took a large variety of configurations, and where the trainees were randomly assigned to configurations. Such data did not exist and obtaining it would have been extremely difficult and time-consuming. Rather than trying to conduct such research, it was decided to collect and analyze existing data to see if a summary of experience could be assembled. This approach cannot rigorously support inferences of causation, but it can produce baseline data. Such information may be used cautiously for decision-making and can be used as a guide to further data collection. In particular, the simulation developed in the present study provides the basis for organizing future research on the effectiveness of training in producing TOEIC score change.

Examinee records in the data sets used in this study included pre- and post-training TOEIC scores, known lengths of training, and descriptions of the training in terms that could apply to the different data sets. However, information on course conditions was not available for some of the cases. Rather than eliminate these cases it was assumed that the training followed the frequently noted combination of having the purpose of general education, using general educational materials, requiring a bachelor's degree but not

specifying the major, and having moderate class sizes. It was assumed that specialist training for instructors, special materials, or a particular focus of the training would have been mentioned in the records of the training or known by the companies where the training took place.

The general analysis plan of this study was to use pre-training test scores, time, and coded variables in regressions. The coefficients of these regressions are given in Table 6. Because data on a complete set of all the combinations of course conditions were not available, the analysis was limited to main effects only. For example, an effect of using a BusSim was calculated without regard to whether that simulation was used by a specially trained instructor or for what purpose it was used. The use of the constants in Table 6 to calculate predicted post-training TOEIC scores demonstrates what is meant by not considering the possible facilitating effects of course conditions in particular combinations; Table 6 contains no line items with combinations of course conditions. As a result, the findings presented here are somewhat limited. It may well be that certain course conditions have considerably greater (or lesser) effects under specific conditions. Future studies are required to learn about possible facilitating effects of course conditions in particular combinations, where the existence of such effects are suspected.

The simulation shows that the class purpose, materials, and instructor preparation were major factors in determining the number and post-training TOEIC scores of trainees reaching or exceeding a specified proficiency level. It should be understood, however, that these results apply to the types of treatments used in the study. Each course was, in the opinion of the trainers, a credible process for preparing trainees. For example, in the present study the effect of varying total time was not large. But one must keep in mind that the courses used in this study were of sufficient length in the opinion of the trainers that the material could be covered at a reasonable pace; the results do not imply that courses could be shortened to the point where the material to be taught cannot be covered at a reasonable pace. Also, in order to obtain the gains observed due to using Video in this study, the Video must be used appropriately and incorporated into the class structure, as was undoubtedly the case in the courses whose data went into this study. The simplicity with which the arithmetical manipulations simulate changes in course conditions should not lead one to overlook the fact that, as always, courses should be carefully planned. Ideally, descriptions of the courses used here would have been appropriately detailed. This was not possible, hence the study puts a heavy burden of reasonable judgement on the user of the results. It is recommended that a careful description of the conditions of future research and training be written and kept on permanent file.

The original interest in this study was driven at least in part by cost. The simulation can help in cost planning in that costs can be associated with the elements that go into the simulation. The results can provide information on simulated training yields, which can help put the costs on a dollar per qualified trainee basis. For example, if one is considering introducing the use of newspapers/ current events into a course that currently just uses general English materials, a cost analysis could establish fixed costs associated with installing the new materials, and per-trainee costs associated with completing the training of one trainee regardless of the outcome of the training. The simulation could then be used to estimate the training yield afforded by the new course. High training yields would keep the cost per successful candidate low. The training yield could be used to calculate a cost per successful candidate or per fixed number of trainees reaching a specified post-training proficiency level. In carrying out such calculations one should be sure that the TOEIC score distribution used to simulate input to training is representative of those actually being considered.

One procedure for showing how the simulation could enter into cost considerations would be to compare installing Video training with just raising the cut score for selection into training. Raising the cut score would reduce the training cost per successful candidate at a very small installation cost, because the pass rate would increase. However, the number of trainees successfully completing the course would decrease because fewer candidates would qualify for training.

So long as there is a substantial pool of able candidates for training, raising the cut score is probably quite cost effective.

However, in a personnel shortage situation the goal would be to increase the training yield drawn from an already inadequate pool of candidates. One could consider the costs and yields of several types of training and make an informed decision as to how to meet the resulting manpower requirement. One might also let the simulation determine what combination of costs and treatment effects would lead to an adequate yield of personnel, using the result of the simulation as a conceptual background for designing and introducing a new type of training.

## Concluding Comments

This study helps to define the relative contribution of several training factors to student improvement in English language proficiency. However, given that language learning is affected by many more variables than just those examined here, the program administrator and teacher ultimately will need to examine the data and make their own judgments as to the relative efficacy of various approaches to training.

This study also points to several directions for further research. In this study, course materials were frequently confounded with teaching methods. For example, business simulations involve both materials (e.g., real-life case studies, informational material) and teaching techniques (e.g., role play). Although it is often difficult to separate materials from methods, future researchers may be able to create different ways of categorizing these variables so that we can more adequately address the differential impact of methods and materials. In addition, other factors that we know impact learning and teaching, such as student motivation, should be included in studies to more realistically evaluate the influence of training in a classroom setting.

The current study, while faced with several unavoidable limitations, does help to shed light on an important area that holds great promise for language educators everywhere. These results should be taken as a first step towards a better and more complete understanding of the effects of language training on student progress.

## References

Cohen, J., & Cohen, P. (1983). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Mawah, NJ: Lawrence Erlbaum.

Dobson, A. J. (1983). Introduction to Statistical Modelling. New York: Chapman and Hall.

Hardy, M.A. (1993). Regression with Dummy Variables. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-093. Newbury Park, CA: Sage.

Huitema, B. (1982). Analysis of Covariance and Alternatives. New York: John Wiley and Sons.

Kempthorne, O. (1952). The Design and Analysis of Experiments. New York: John Wiley and Sons.

Lord, F.M. (1960). An empirical study of the normality and independence of errors of measurement in true scores. Psychometrika, 25, 91-104.

Mollenkopf, W. (1949). Variation of the standard error of measurement. Psychometrika, 14, 189-229.

Saegusa, Y. (1985). Prediction of English proficiency progress. Musachino Women's College English Literature Society, 18, 165-185.

Snedecor, G.W., & Cochran, W. G. (1979). Statistical Methods. Ames, IA.: Iowa State University Press.

Stuart, A. (1991). Kendall's Advanced Theory of Statistics. New York: Oxford University Press.

Wolf, F. M. (1986). Meta-analysis: Quantitative Methods for Research Synthesis. Beverly Hills, CA: Sage.

**Pages 26 and 28 are blank and have been deleted from this PDF.**

# APPENDICES

**A. Non-linear prediction**

**B. Dummy variables**

**C. Linearity and homoscedasticity**

**D. Skewness and kurtosis**

## Appendix A. Non-linear prediction

Non-linear prediction formulas were developed in the estimation sample using neural nets (Caudill, 1990) with one hidden layer and five nodes. The nets used both pre-training Listening and Reading scores, all three time variables, and allowed an adjustment for each company supplying data. The output of the nodes in the hidden layer were logarithmic transformations of their input. The computational algorithm used to estimate the constants in the net was a multivariate Newton iteration used to minimize the least square fit of predicted post-training score to observed post-training score. Exploratory analyses suggested that using more than five nodes added to the computation time but gained little in terms of predictive validity in the estimation sample, and lead to greater shrinkage in validity in the cross sample.

## Appendix B. Dummy variables

The imputations discussed in this report were necessary to estimate and test course condition effects. In addition, a change in the statistical models used for Tables 1-3 was necessary. Those models indexed which participating company provided data for a case by using dummy variables. Each company used one, or at most a few, sets of combinations of course conditions under which the cases from the company were grouped. It was necessary that the dummy variables index the course conditions that were used in the training, rather than the company. This was done by the following procedure. After the first few columns of the data matrix were allocated to test and training time variables, the remaining columns were assigned to course conditions. For example, in the column to which the Video condition was assigned a case would be assigned a 1 if the course used Video and a zero if it did not. Similarly, other course conditions were assigned to columns that contained zeros and ones according to the conditions under which the particular cases (examinees) were trained. In the data matrix a row represents an examinee, so rows and columns of the data matrix contain a pattern of zeros and ones that represent the distribution of course conditions to cases. The pattern of zeros and ones in a row indexed the course conditions used by the company employing the examinee represented by the row.

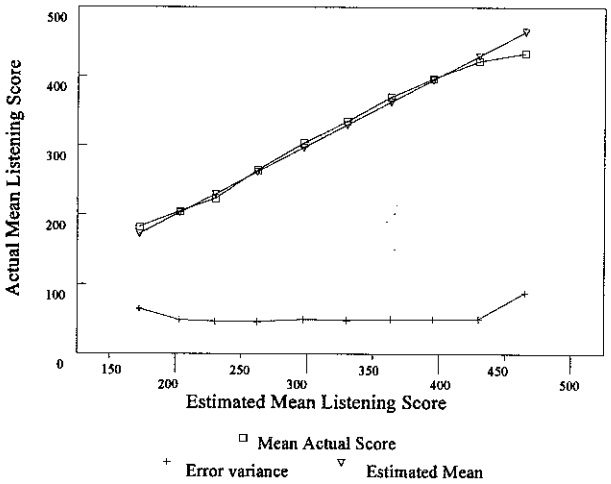
When the data matrix was set up in the manner described above, multiple regression computations were used to estimate the treatment effects. The computation resulted in quantities that took effect or not according to whether the dummy variables were zero or one. However, a limitation of this procedure is that the columns of the data matrix must be linearly independent. To achieve this independence when such a system is used with analysis of variance or covariance, treatment means are usually estimated as deviations from a grand mean which is taken as being present for every case. When taken as deviations, the treatment effects sum to zero. Hence, one treatment effect can be dropped from the data matrix (its value being the negative of the sum of the others).

However, in the present analysis, rather than using the grand mean we used a basic combination of conditions as a base-line effect to which the other effects were added. Course conditions making up the basic combination included the use of general English materials, a BA required for the instructor, general education as the purpose of the course, and a class size of between 10 and 20 students. The value of the dummy variable for the basic combination was set at one for every case, with others set at one if the actual condition for the case was different. For example, to calculate the combined effects of course descriptors for case A who was trained using Video for staff development purposes in a class of moderate size, we added to the weight for the basic combination the weight for Video plus the weight for staff development, i.e. the dummy variables for the basic combination, Video, and class development would be one and all other dummy variables would be zero. No adjustment for class size was needed because case A was trained in a moderate sized class and because that condition was part of the basic combination. Had case A been a new employee, then an adjustment using the weight for new employees would have been necessary. For further discussion of dummy variables see Hardy (1993) and Kempthorne (1952).

## Appendix C. Linearity and Homoscedasticity

Figure A presents results on linearity and homoscedasticity in the Listening data. For this Figure, predicted post-training Listening scores were calculated for all cases in the cross sample. The predictions were divided into 10 equal intervals and, for each range, predicted and actual mean scores were computed, as well as the variance of the errors of prediction. The means actual scores and variances of the errors of prediction were plotted against the mean predicted scores for each range. In addition, the mean predicted scores were plotted against themselves to provide a linear reference line with which to compare the plot of the mean actual scores.

**Figure A:** Listening Comprehension Mean and Variance for Estimate Score Levels




Examination of Figure A reveals that the plot of the actual means was almost collinear with the plot of estimated means. This result strongly supports the use of the linear system that was chosen in this study. Only at the extreme top and bottom of the range does the plot of actuals against estimates flatten slightly. Note also that the plot of variances of errors of prediction against estimated scores was virtually flat and horizontal. Again, only at the extremes of the range of estimated scores does the plot of variances curve upward, and the upward curve at the lower end was extremely small. This plot demonstrates that, relative to variation in the predicted and actual scores, the variances of the errors of estimate were homogeneous.

Figure B presents results on linearity and homoscedasticity in the Reading data. This plot was constructed in the same manner as was Figure A, with the exception that the Reading cross sample data were used. Again the plots reflect linearity and homoscedasticity throughout most of the range of predicted scores with slight departures at the extremes. Departures from homoscedasticity at the extremes were also slight.

**Figure B:** Reading Comprehension Mean and Variance for Estimated Score Levels

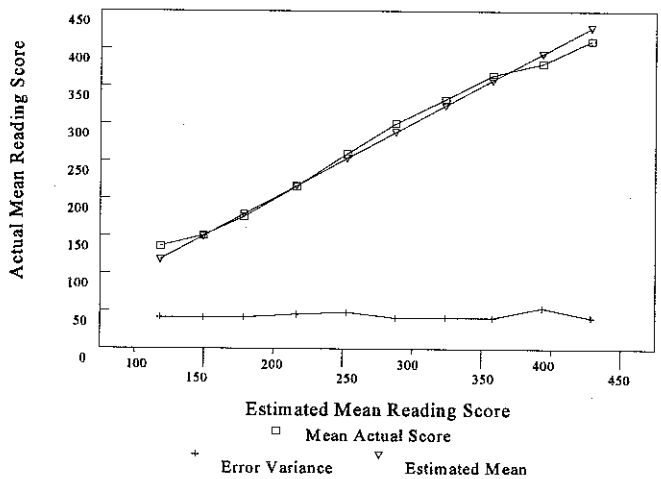
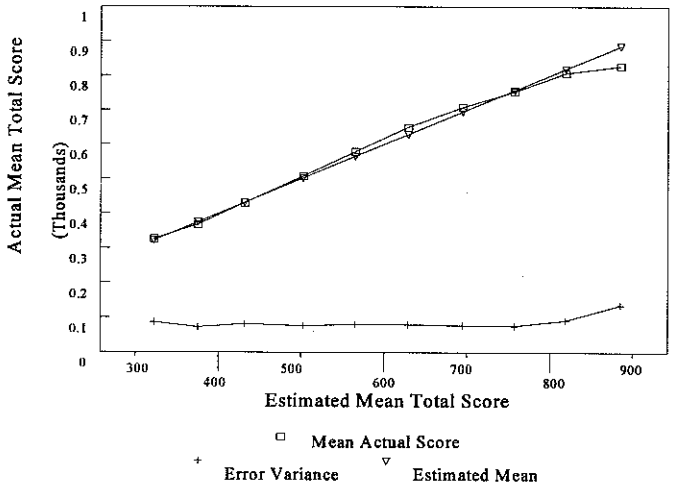



Figure C presents results on linearity and homoscedasticity for the Total score data. This plot was constructed in the same manner as were Figures A and B, with the exception that the Total score was used from the cross sample. As with Figures A and B, the plot of actuals against estimates flattens slightly at the extremes, and there is a slight upward curve of the variance at the extremes.

It was not absolutely necessary that the errors of estimate of the post-training TOEIC scores be homoscedastic in the cross sample, but the simulation was much simpler to operate than if the errors had been in some way dependent on the

predicted score. Indeed, there was reason to expect such a dependency. Mollenkopf (1949) and Lord (1960) have demonstrated that the variance of errors of measurement are larger in the center of the test score distribution. Mollenkopf (1949 and in Gulliksen, 1987) also gave a theoretical rationale for this relationship. Mollenkopf's findings and derivation apply to the present situation in that they deal with scores on pairs of parallel tests, and the forms of the TOEIC test are regarded as parallel. However, in Mollenkopf's study the parallel tests were administered on a single occasion whereas in the present study they were administered on two separate occasions with training during the intervening time interval. Perhaps the intervening events blurred these effects in the present study. In any case, the burden of proof was on the present study to demonstrate that homoscedasticity obtained, and fortunately it did.

**Figure C:** Total Score Mean and Variance for Estimated Score Levels



## Appendix D. Skewness and Kurtosis

The index used for skewness was the ratio of the average cube of the errors of prediction divided by the cube of the standard deviation; the index for kurtosis was the average fourth power of the errors of prediction divided by the squared variance. In the normal distribution, the skewness index would be zero and kurtosis index would be 3. Standard errors of these coefficients were available in Snedecor and Cochran (1979), hence the ratio of the index of skewness to its standard error could be used to calculate *z*-scores that could be referred to normal tables to estimate significance levels. A similar process yields the significance level of the coefficient of kurtosis if 3 is subtracted from that coefficient before the *z*-score is calculated. The results of these tests appear in Table A.

Note in Table A that only the coefficient of skewness for the Listening errors of estimate conformed to the value required for accepting the hypothesis of normality. Hence the normal distribution was not adopted for use with the errors of prediction.

**Table A.** Estimates and tests of significance of coefficients of skewness and for Listening, Reading, and Total errors of prediction

| | Listening | Reading | Total |
|---|---|---|---|
| ***Skewness*** | | | |
| Coefficient | -.059 | .308 | .304 |
| *z*-score | -1.279 | 6.692 | 6.600 |
| *p*-value | .20 | <.0001 | <.0001 |
| ***Kurtosis*** | | | |
| Coefficient | 6.232 | 4.445 | 5.884 |
| *z*-score | 35.099 | 15.695 | 31.321 |
| *p*-value | <.0001 | <.0001 | <.0001 |

## TOEIC Research Publications

### *TOEIC Research Reports*

Wilson, K.M. (1989). Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experience in the TOEIC testing context, TOEIC Research Report No. 1 (99 pages).

Dudley-Evans, R., & St. John, M.J. (1996). Report on Business English: A review of research and published teaching materials, TOEIC Research Report No. 2 (45 pages).

Boldt, R.F., & Ross, S. (1998). Scores on the TOEIC (Test of English for International Communication) test as a function of training time and type, TOEIC Research Report No. 3.

### *TOEIC Research Summaries*

Woodford, P.E. (1982). An introduction to TOEIC: The initial validity study, TOEIC Research Summary (16 pages).

Wilson, K. (1993). Relating TOEIC scores to oral proficiency interview ratings, TOEIC Research Summary No. 1 (11 pages).

Boldt, R.F., & Ross, S. (1998). The impact of training type and time on TOEIC scores, TOEIC Research Summary No. 3 (10 pages).

** Please note that there is no TOEIC Research Summary No. 2.*

All TOEIC research publications are available for a small fee (plus shipping and handling) from your local TOEIC representative, from The Chauncey Group Europe SA in Paris, France, or from The Chauncey Group International in Princeton, New Jersey, USA.