TOEIC Research Report

Language Proficiency Gain on the Test of English for International Communication: Meta-Analyses of Japanese and Korean Corporate Language Programs

R. F. Boldt and S. J. Ross
Educational Testing Service TOEIC Technical Panel &
Princeton, N.J. School of Policy Studies, Kwansei Gakuin

Abstract

The globalization of industrialized economies has increasingly required corporations to train employees for overseas posting. It has also involved the need to conduct in-house foreign language training. Such programs are frequently evaluated for cost effectiveness, leading program designers to search for methods of providing evidence of language gain accruing from instruction. Accountability in such programs is often contingent on a comparison of standardized test scores before and after the program of instruction. A parallel need in language program evaluation is in describing not only that there has been language gain, but also in providing information about features of the program that are related to gains. An inferential problem with pre-and-post comparisons comes when programs are evaluated without reference to comparable controls. One methodological alternative to conventional language program evaluation is a meta-analysis of cumulative archives of records gathered from a variety of programs. The present meta-analysis of language programs is one such example. Tests and surveys from 36 language programs in Japan and Korea provided descriptors of course objectives, language curriculum features, materials, duration of instruction, information about instructors, and pre-and-post instruction TOEIC scores. In all, more than 3200 records of pre-and-post test scores were coded for salient program characteristics. These characteristics were used in 'dummy codes' for input to multiple regression analyses devised to estimate effect sizes for different combinations of program characteristics. The results of the meta-analyses of aggregated data describe salient patterns of teacher training, effective material use, and program duration that are linked to benchmark TOEIC outcomes in corporate programs. The outcomes of this study extend to implications for optimal language program organization.

Second language program evaluation has in recent years included an increasing variety of analytical methods. These methods range from experimental or quasi-experimental to ethnographic (Alderson and Beretta, 1992; Lynch, 1996). With such variation in research methods, it has become considerably more difficult to summarize second language program evaluations in a comprehensive manner that can provide aggregate effect size estimations. Despite this difficulty, there continues to be a need for general summative overviews of language education program outcomes. One motivation for establishing a general overview of the product of second language program evaluations is that it potentially provides an aid to program developers and curriculum designers. Given the wide variety of instructional options available, task-based, forms-focussed, natural approach, to name but a few, and rapid changes in instructional technology, characteristic features of programs that have shown patterns of success may serve to provide a proactive strategy for language program designers.

The present project aims to provide one such general survey of outcomes in the form of a meta-analysis. The approach taken here is to extract program characteristics from a wide variety of programs in Japan and Korea which have used a single criterion measure for program evaluation, the Test of English for International Communication (TOEIC). The advantage of conducting the meta-analysis using a single criterion measure comes from the comparability of the scale defining gain across different studies. In conventional meta-analysis (Wolf, 1986; Hunter and Schmidt, 1990), different tests of main effects, sometimes using strikingly different criterion measures, need to be standardized before estimation of mean effect across the studies. With different criterion measures with varying degrees of internal consistency, conventional meta-analysis runs the risk of estimation error. The use of a single criterion method in the present study reduces this

risk. This is accomplished by pooling data sets instead of the conventional pooling and averaging of effect sizes from statistical tests of the observed main effects across studies.

A similarity between the meta-analysis presented here and conventional meta-analyses exists in the fundamental effort needed in coding program characteristics (independent variables). Traditional surveys of published research outcomes need to identify independent variables across studies as indicators of common underlying hypothesis. The meta-analysis presented here derives the coding ex post facto through the use of surveys of program administrators. Instead of a meta-analysis interpreting independent variables in relation to criterion main effect tests, the approach taken here was to survey administrators using a standardized survey instrument. This approach, merged with the common criterion measure in a pre-post design, provides the potential for a meta-analysis with less interpretive variation than that required with the use of published research outcomes.

Program Variables

In the survey summarized here, key characteristics of corporate language teaching programs were surveyed. Input to the survey came from interviews with several corporate language program directors in Japan. The initial interviews provided us with key parameters about language program organizational features, which could be built into a standard survey. The final version of the survey was sent to program managers at a wide range of company programs in Japan and Korea.

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 are designed to feature language in the context of its use. Recent innovations in communicative materials now often include the use of language learning tasks, language functions, video, simulation tasks, and role plays. The communicative approach contrasts with the structuralism characteristic of the 1960'-70's which tended to feature language structure atomized into constituent parts. The focus then was to build proficiency from the bottom up through the use of sequential structural analysis in a "building block" manner.

The materials sampled in the current meta-analysis were in large part communicative. They differ mainly in the scope and focus in featuring English language in different contexts of usage. Some, for instance, may feature language use in the context of on-the-job interaction. 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. Other materials were not contextualized in any given job-specific format. These may rely more on learners' world knowledge in the most general sense. The focus for this genre of materials is to help students associate the form of language with their own communicative intentions about issues anticipated to be of practical interest.

The classification of materials was based on two major criteria. The first was the client company's identification of the type of materials used in the particular courses featured in the meta-analysis. The second was, in some cases, a corroborating examination of the actual course material specified by the client company. This two-tier system led to a general classification system of the teaching material. It should be noted, however, that some classifications are not mutually exclusive. At some client companies a heterogeneous

approach to materials selection was used, and therefore more than one type of materials may be indicated. The emphasis in the classification exercise was to identify the predominant type of materials used in the featured courses.

The listings below outline the major features of the materials used. The materials are described in terms of primary language syllabus organization principles, and in terms of their potential overlap with the TOEIC test content domain.

Business Simulation (BusSim). A popular approach to language pedagogy for company employees is based on the process of gaming or simulation. For this, materials are devised to match a context of use in which the language learners are most likely to use the target language. Simulations may vary by job specification. Business simulations often present role play scenarios in which learners attempt to use the target language in a plausible 'real life' context. An example might be dealing with an overseas client who has sent a fax message complaining of the non-delivery of spare parts. The learners then simulate an interactive telephone call in which they act out the role of the overseas client and the company employee nominated to explain the details of the shipping problem to that client.

Business simulations often follow a deductive learning scheme. Example contexts and interactions are presented on the form of dialogues, which are then practiced. Interaction rules or patterns are learned, often as formulas, and then practiced intensively. The role plays then follow and are repeated until the learners are reasonably confident that they could apply the formulas in comparable contexts in the work place. Such simulations are often devised to be 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. The language featured in business simulations usually does not involve careful

linguistic analysis of specific forms, as would be done in a structural syllabus. The major connection between business simulation materials and the TOEIC is in the domain of 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 rarely in the formal characteristics of language for business.

General English (GenEng). In the context of the meta-analysis, the classification of materials as 'General English' is the most variable. General English texts differ considerably in quality and domain sampling. Most texts of this genre are published by international publishing concerns and are marketed globally. Most of these texts are designed to be used 'off the shelf' by minimally trained native-speakers. A few require in-house training seminars devised 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 everyday scenarios for promising or making an appointment may be contrasted in the lesson segment. In this regard, the 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 narrowly sample language functions in the context of business interactions. General English texts usually feature some form of language structure analysis as well. 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 weight the listening and speaking skills more heavily, because these are the two skills that are the most underdeveloped by the Japanese educational system in general.

Compared to business simulation materials, general English materials seem to better match the domain of language sampled by the TOEIC test. This coverage is mainly in the sampling of 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 replacing a printer cartridge), a wider sampling of language structures, and a 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). An unstructured approach to language teaching material features current news items 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. The primary difference is in language selection. Newspapers and current events magazines are not edited for non-native readers, and thus may feature a wider range of idiomatic and specialized usage not common to international English, or to business domains of language use. No interface to language functions is provided in news materials.

The usual approach to using current events materials is in having learners read and then discuss opinions about the news. In this regard, the main purpose of these materials is more related to language maintenance and practice than to learning of new structures or

functions. The use of current event materials can nonetheless lead to incidental learning of new vocabulary and grammatical forms.

The interface of current events and news materials to the content of the TOEIC tests is considerable for one dimension of the test. The wide range of language forms appearing in non-pedagogical texts sample a comparable domain of usage seen on the TOEIC reading/structure section. The major difference is perhaps on the listening skill. Since the input from current events materials is primarily visual, learners' experience with listening tasks like those featured on the TOEIC test is probably most 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). An 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 'original' samples of authentic language for non-pedagogical purposes, the salience of form and function can be thought to differ. Materials designed to teach forms and functions make these features of language optimally salient though the scripted interactions of the actors. Authentic materials, which are more common, make no such direct attempt.

Video materials mainly differ from the current events and news materials in the input channel aspect. Current events and news are usually presented in the written input mode. 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. We can anticipate, therefore, that there would be a differential interface between these two types of materials with the content and format of the TOEIC test.

The video materials probably 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; many if not most Japanese learners have greater reading and grammar skill development than aural comprehension at the beginning of company program instruction. The choice to enhance what is least developed is often a pragmatic one, considering company needs for employees who can use English for oral communication.

The interface of video materials to the TOEIC test appears to be the opposite of that for the current events and newspaper materials. If the input is mainly audio-visual, we would anticipate that language development would be largest in the area of aural comprehension. The listening comprehension portion of the TOEIC test could therefore be anticipated to differentially assess gain accruing from the use of video materials.

Instructor Variables

Bachelor's degree (BA). Many company programs and language schools employ young college graduates as 'native 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 often not considered important as a hiring criterion in company programs or language schools. Teachers with the minimal qualification were encoded 'B.A.' in the meta-analysis.

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, usually devised on an in-house basis. The training regimen rarely includes an outline of theories of cognitive development, language acquisition, or assessment, for their chief purpose is to ensure some uniformity in the use of the materials sold as part of the language training package. Programs providing in-house training for teachers were encoded 'PRG' in the meta-analysis.

Certificate. Language teaching certification courses such as the Royal Society for the Arts TEFL Certificate are not uncommon for expatriate language teachers. These courses usually provide an orientation to practical teaching methodology for in-service teachers of English. While the range of the curriculum is certainly much broader than 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. Instructors with this kind of training were encoded with 'CRT' in the meta-analysis.

Master degree (MA). Instructors in language schools and company language training departments holding masters degrees in any field are still relatively rare. In the context of the present study, any master degree achieved was coded as 'M.A.'. The possible ambiguity in encoding a bachelor degree in any field as a single label also applies to the coding of master's degrees. Some bachelor degrees, for instance, may be in a modern language or linguistics, while some master degrees may in fact be in diverse fields ranging from finance to 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 effect sizes attributable to educational qualifications.

Objectives

Company programs provide language training programs to promote English language development for employees showing potential for overseas posting. In some programs, employees with needed technical expertise are given language training regardless of their aptitude or prior experience in language learning. This phenomenon is related to the increased globalization of off-shore production by Japanese corporations. In contrast to the need for overseas sales staff a decade ago, the current trend has been to train production line personnel for overseas posting.

Language schools and college programs, in contrast, are typically less influenced by specific outcome expectations. The focus in these contexts is on the development of overall language proficiency, and often just an appreciation of foreign language literature in translation. The coding system used in this study therefore relates to the stated goal of the program as specified by the program administrators.

General education (GenEd). The most common objective was for general language proficiency development - encoded in the meta-analysis as 'GenEd'. This goal corresponds approximately to the use of four-skills texts, and occurred in all of the types of programs sampled in the survey.

Staff Development (StaffDev). Company programs differ from language schools or colleges in that they may 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 employees even without specific plans to post them overseas or give them job assignments requiring proficiency in English. This type of objective was coded 'StaffDev' in the meta-analysis, for it identifies the policy of the company program to create

a pool of employees from which the company may draw persons with English language skills in order to meet future company needs.

New Employees (NewEmp). Corporate programs differ in another aspect. New employees may be assessed with TOEIC in order to index their aptitude for further language training. Once identified, new employees may be given intensive language training in-house. This strategy, labeled 'NewEmploy' in the meta-analysis, indicates a company policy of employee evaluation prior to more permanent job assignments in Japan or overseas. In contrast with the staff development objective, the new employee training objective is seen as pre-posting evaluation of employee's readiness for different kinds of jobs required in the corporation.

Class Size

Organization varied considerably in the programs sampled in the facet of class size. This facet has a clear potential impact on the rate of gain in language development, since the amount and focus of input and feedback 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 on TOEIC, the following codes were used: Classes larger than 20 students were labeled "Large"; Those with more than ten, but less than 20 were called "Mid"; classes with fewer than 10 students were considered "Small" classes in the meta-analysis.

Class sizes corresponded to program types. College programs typically have very large classes (35 students per class). Company and language school programs are usually in the middle to small class range depending on the program objectives. Extensive courses, in which instruction given over several months, are usually middle-sized. Intensive courses tend to have a small teacher to student ratio

Data Sets

Data sets were obtained from 23 Japanese and 13 Korean companies and language training institutions. These institutions supplied pre-post TOEIC data on from 30 to 1,030 trainees. The language proficiency test data supplied were Reading and Listening TOEIC scores, except in one case where only TOEIC total scores were available. Training time information comprised total number of weeks and total number of hours. In most of the analyses to be reported the training time figures used were those intended for the course and did not include actually attendance times. Other types of information on the training included the treatments as recounted above. For each type of information, only one category was reported for each subject. For example, even though both video materials and newspapers were used in some course, only the predominant material was identified, e.g., video or newspapers but not both.

The Japanese data did not include complete information instructional materials. As a result there was a great deal of variation in the percent of cases for which the information on the variables mentioned above was missing. Further, because some time had passed since data collection actually occurred it was not feasible to obtain all of the missing information. The amount of missing data was an important factor in determining the major analysis groups in the Japanese data. The major analysis of the Japanese data was conducted using the 4,247 cases for which all of the Listening score, the Reading score, the number of weeks of training, the total hours of training, and for which the class size was larger than ten. This sample will be referred to as the "major Japanese sample." The major analysis of the Korean data was conducted using 1627 cases for which the same variables were present as were required for the Japanese data.

Methods of Analysis

The major analyses in the study included two steps: (a) estimation of constants in such as regression weights or treatment effect sizes, and (b) evaluating prediction results obtained by prediction functions that used the constants estimated in step (a). Steps (a) and (b) used comparable data sets but did not contain the same cases. Rather, for both the Japanese and the Korean data sets the major samples were divided into two sub-samples. The first sub-sample was used for estimating the constants and is referred to as the "estimation" sample; the second was used for evaluating prediction results and is referred to as the "cross-validation" sample. The division of the major sample was accomplished by removing every third case, resulting in two sub-samples, the estimation sample having 1416 cases and the cross sample having 2831 cases. When creating the Korean samples, the estimation and cross samples each used every other case, resulting in an estimation sample of 814 cases and a cross-validation sample of 813 cases.

Non-linear prediction formulas were initially developed using neural nets (Caudill, 1990). The neural nets used to obtain the above results had one hidden layer and five nodes. The outputs of the nodes in the hidden layer were logarithmic transformations of their input. The computational algorithm used to estimate the constants used 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. This analysis was done in the estimation sample. Exploratory analyses suggested that using more than five nodes adds to the computation time but gains little in terms of predictive validity in an estimation sample, and leads to greater shrink in validity in the cross sample. The nets used both pre-training Listening and Reading scores, all three time variables, and allowed an adjustment for each company supplying data.

It was found that linear and non-neural net predictions predicted equally well. The resulting multiple squared correlations (R^2s) were .65 for Listen, .74 for Read, and .74 for Total, and these same values obtained whether the prediction formula was linear or neural net if all predictors were used. The linear prediction system is easier to understand, easier to use, and less prone to validity shrink in a new sample. For these reasons, the use of non-linear prediction was not considered further.

For the Japanese and Korean data, a separate multiple linear regression model was used to develop systems for predicting post-training test scores. The purpose of the regression analyses was to develop [a] weighted composites of predictor scores and [b] sets of adjustments to the composites to account for course variables. The weights and adjustments in the adjusted composites were chosen so as to maximize their correlations with post-training scores. These analyses used the data in the Japanese and Korean estimation samples.

Imputation criteria and methods: The computations described above are consistent with the analysis of covariance model. However, because no record or memory of course conditions were available for a large number of cases in the Japanese data it was necessary to impute conditions to replace the missing data. Imputations used were as follows: "general English" was used where the course material was missing; "BA" was used when the instructor requirement was missing; "MId" was used when the class size was missing. The purpose of the training, e.g., GenEd, was indicated for all cases. This strategy differs from Bayesian methods (Schafer, 1997) in that the imputations are based on the most frequent observed category for each variable.

The rationale for imputing general English as the course material when there was no actual record of that material was as follows. Other possible imputations were News, Video, and BusSim. Though one company used newspapers, that company was not part of the

major sample. The use of both video and business simulation would require substantial preparation, which it was felt would be remembered by our informants even though some time had passed since the courses were delivered. 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.

In the Japanese data the complete range of qualifications of the instructors included general Bachelor of Arts, any Master's degree, certificate in teaching, or training in specific teaching materials. Of these, only BA and training in specific teaching materials occurred in the major sample. Since the training in specific teaching materials would require substantial preparation and would be remembered by our informants, it seemed prudent to impute BA as the requirement when the requirement was not known.

No imputation was necessary for the purposes of the courses, but not all of the class sizes were known. For only two companies were class sizes known to be less than 10 or greater than 20; all the rest of the class sizes were known to be intermediate or were unknown. In this case, it was considered safest to impute an intermediate size to classes of unknown size.

Information on the training conditions affecting the Korean data was far more complete than for the Japanese data. Hence, No special imputation procedures were used for the Korean data.

With the imputations described above, the Japanese sample required a complete 3x2x3x2 complete factorial design. BusSim, GenEng and Video were the levels of the first factor, which was the materials factor. The codes for BA and PRG were the levels of the second factor, which was the instructor requirement. The course purpose factor included levels for GenEd, StaffDev. Mid and Large made up the fourth factor, which was class size.

The Korean sample took a 5x3x4 complete factorial design as follows: GenEng, TestPrep, Unknown, BusSim, and Conversation were the levels of the first factor, which was the materials factor; TES, BA, and MA were the levels of the second factor, which was the instructor requirement; and GenEd, Overpost, StaffTrain, and Upgrade were the levels of the third factor, which was the purpose of the training. Numbers of students in the Korean classes were unknown. 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. The likelihood ratio criterion test of significance was a feature of the significance tests. This procedure was adopted because of its convenience in sequences of significance tests such as occurred here. It is a large sample test, which is also consistent with the sample sizes in the present research.

Briefly, this procedure involved fitting a complete set of predictors of post-training TOEIC performance, then fitting all but a given factor of interest, and then comparing the two using a chi-square with degrees of freedom equal to the number of prediction weights estimated. The complete set of predictors comprised TOEIC scores, the three conditions of training, number of weeks, total training time, and intensity (total time divided by weeks). For example, when testing to see if equal weights could be used for the two TOEIC scores, Listen and Read, the complete set of predictors included the list given above and featured two regression weights, one for Listen and one for Read. However, for the second fit the weights for Listen and Read were made equal, leaving one less regression weight to be determined, hence requiring a chi-square with one degree of freedom. The likelihood ratio criterion procedure (Dobson, 1983; Stuart, 1991) was used to compare results. When used in the present context the likelihood ratio criterion procedure yields a chi-square with degrees of freedom equal to the difference in the number of predictors used by the results

compared. The regression approach follows a step-down sequence with the elimination of individual variables and monitoring of changes in the coefficient of determination.

Results

The results of the analysis are framed as program planning issues likely to be considered important by language program designers.

Program Issue 1: Should a training program be intensive or extensive?

In order to examine this issue, we model pretest TOEIC total scores, hours of classroom instruction, weeks of instruction, and the intensity of instruction calculated as hours per week.

Table 1j Composite prediction of TOEIC total test score (Japan)

Predictor Set	R^2	Contrast	Chi-square
1) All Variables	.7281		
2) drop Wks,Int	.7281	1 vs 2	n.s.
3) drop Hrs,Wks,Int	.7268	1 vs 3	p <.02

All Variables = L1,R1,Weeks,Hours,Intensity. N = 1416

Discussion of 1j. It appears for the Japanese data set that the intensity of instruction does not have a significant impact on gains. As in other research on language program impacts (Saegusa, 1986; Ross, 2000), the duration of language study (hours) is the major controlling influence on the rate of improvement. It appears that organizing the training program to be highly intensive does not lead to a better outcome than a longer, less intensive program.

Program Issue 2: What impact does instructor qualification have on outcomes?

Table 1k Composite prediction of TOEIC total test score (Korea)

Predictor Set	R^2	Contrast	Chi-square
1) All Variables	.8708		
2) drop Inst. Qual	.8691	1 vs 2	$p<.05$
3) drop Weeks	.8705	1 vs 3	n.s.
4) drop Hours	.8707	1 vs 4	n.s.
5) drop Intensity	.8706	1 vs 5	n.s.

All Variables = L1,R1,Instructor Qualification, Weeks, Hours, Intensity. N = 1626

Discussion of 1k: Holding pretest listening, reading scores, and instruction time constant, it appears that instructor qualification has a significant impact on gain, especially in the Korean context. The inference we make here is that more qualified language teachers (M.A./M.Ed. degree or CRT Certificate, versus B.A./B.S degree) can better influence the learning environment. We assume that this advantage accrues from M.A./M.Ed qualified teachers are more likely to have had exposure to TEFL methodology courses, which eventually equip them better for classroom teaching. We will return to this issue later when we conduct simulations.

The original research interest in this study was driven at least in part by cost management. The simulation can help in cost planning to the extent that costs are associated with the elements used in the simulation. The simulation provides information about training yields, which can help put the costs on a per-trainee basis. For example, if one is considering introducing the use of video materials into a course that currently just uses general English materials, a cost-effectiveness analysis (Levine, 1983) could establish fixed costs associated with installing the new materials, and per-trainee costs established 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 available to post-training assignment. In carrying out such calculations one should be sure that the TOEIC score distribution used to simulate input to training be representative of those actually being considered.

Program Issue 3:What is the most typical type of program?

The most typical set of characteristics of programs are derived from the modes of variables gathered in the survey. From these modes, a baseline program can be described and its impact can be modeled using the pre-and-post test TOEIC scores. In addition, we are interested in modeling two ranges of scores that can be interpreted as ranges likely to be considered in program evaluations. The 'lower bound' range is denotes a score range that would be characteristic of employee eligibility for an in-house language program. The survey revealed that a minimum of 220 would be a typically lowest score. The exit score (deemed the 'passing' score) denoting the end of a foundational English course was estimated to be 447. An 'upper bound' range was also derived from the survey, but this range differs in that it represents the range typical of language training for overseas post eligibility. This range covers TOEIC total scores of 651 at entry through 751 as the exit/pass. We will first examine the lower bound baseline derived from the Japanese data.

Table 2a Simulations of Baseline Program Effects

Baseline	* General EFL skills materials are used * Instructor has a B.A. in any field * General Education is the training objective * Class size is from 10 to 20 students

The program impact with these four baseline conditions was simulated by re-sampling the database 500 times and then averaging the means. The percent of candidates surpassing the exit/pass score is then taken as the expected outcome for the baseline condition. In this phase of the simulation, the mean of the baseline was 597.5 for the lower bound, yielding a pass rate of 52% of the candidates.

Program Issue 4: What is more important, teacher training or special materials?

In order to answer this question, we add dummy variables (Hardy, 1993) which encode the use of a particular type of material and then resample the database 500 times. We then average the simulation results and compare these to the baseline mean and pass rate. This procedure is continued by dropping out and adding in different combinations of materials and in-house teacher training characteristics. The results of the lower-bound 'value-added' simulation is shown in Table 2j.

Table 2j Lower-Bound Simulation (Japan)

Conditions	Qualified	Passers	Pass Rate	Mean
Baseline	2633	1365.2	52%	597.5
Business	2633	1636.6	62%	611.3
Video	2633	1746.7	66%	617.0
PRG	2633	1615.7	61%	610.2
Business& PRG	2633	1895.5	72%	624.8
Video& PRG	2633	2003.7	77%	631.3

As can be seen in Table 2j, the effect of adding business simulations to the syllabus is a ten percent increase in the pass rate. By removing the business simulation materials and replacing them with video materials there is a 14% increase in the pass rate - that is, 66% of the program participants would surpass the exit criterion score of TOEIC 447.

The impact of in-house teacher training appears to be less influential than teaching materials, as can be inferred from the 9% increase in the pass rate modeled when PRG (in-house teacher training) is added to the baseline B.A. teacher qualification. The largest value-added increments are seen when teacher training is combined with materials innovation. A 20% increase in the pass rate was modeled when business simulation materials are combined with teacher training done in house. The largest value (a 25% increase over the baseline) is added when video materials and in-house teacher training are used together.

Program Issue 5: What impact does teacher qualification have on outcomes?

Here we take the baseline data as the starting point and find that 64% of the Korean candidates would 'pass' the 447 TOEIC exit criterion (Table 2k). When we model in the effect of teachers having any MA/M.Ed. degree, we notice virtually no change in the pass rate. This implies that unspecified MA/M.Ed. degrees do not translate to more effective instruction or gain on the TOEIC. In contrast, there is a slight increase in the pass rate when teachers have a TESL certification (but not necessarily *any* MA/M.Ed. degree). This outcome corroborates the earlier (Table 1k) observed teacher qualification effect.

Table 2k Lower-Bound Simulation (Korea)

Conditions	Qualified	Passers	Pass Rate	Mean
Baseline	2633	1688	64%	560.6
Any M.A.	2633	1671	63%	558.0
TESL	2633	1766	67%	572.9

Program Issue 6: What best prepares learners for overseas posting?

The candidate qualification range for overseas postings was estimated from the survey data to be 651 as the entry level, and 751 as the exit or 'pass' level. In our data sets, relatively few Japanese and Korean corporations had employees in these ranges (N=372). We nevertheless can simulate the influence of materials and teacher training in same model as that used to address Program Issue 4. Table 3j shows the outcome of the simulation for the Japanese data set.

Table 3j Upper-Bound Simulation (Japan)

Conditions	Qualified	Passers	Pass Rate	Mean
Baseline	372	88.5	24%	810.1
Business	372	144.0	39%	817.2
Video	372	169.1	45%	821.1
PRG	372	139.3	37%	816.6
Business& PRG	372	204.3	55%	827.8
Video& PRG	372	230.4	62%	832.4

The influences of materials and teacher training for the upper-bound simulation parallel those observed in the lower-bound simulation. Business simulation has less impact than video materials for the advanced Japanese learners. As in the lower-bound simulation, the effect of in-house teacher training is not as much of a value-added feature as combining such training with use of business simulation materials (31% increase in the pass rate compared to just 13%) The largest increase is again that modeled when video materials are combined with teacher training, which leads to a 38% increase in the pass rate over the baseline materials. The answer to Program Issue 6 is basically that the same program development strategy used for the lower level courses would work with the clients

qualified for overseas posting - conduct in-house teacher training that focuses on the use of simulation materials or video materials devised for business simulations.

Program Issue 7: Do candidates for overseas posting need specialized teaching?

Here again we conduct the same type of simulation as that done on the lower-bound threshold (Table 2k). Again we find that the difference between a teacher holding any B.A. degree and one holding any unspecified M.A. degree is negligible. As seen in the earlier simulation, the impact of a B.A. plus at least a TESL certification results in a 5% increase in the passing rate at the advanced level. So our answer to Program Issue 7 is a qualified 'yes'.

Table 3k Upper-Bound Simulation (Korea)

Conditions	Qualified	Passers	Pass Rate	Mean
Baseline	372	275	74%	824.7
Any M.A.	372	270	73%	822.3
TESL	372	290	78%	837.2

Conclusions

The first salient point about influences on gain in corporate language programs is that there appears to be a premium associated with internal organization of the program involving active planning and coordination of audio-visual materials or business simulation teaching materials. Additionally, coordination among TESL specialist teachers gives a value-added increment to gains in proficiency (Ross, 2003). The investment in hiring specialized TEFL-qualified teachers is likely to have an eventual effect in better organization of the instruction, which in turn leads to larger program gains on TOEIC. Gain is most clearly associated with the total number of instructional hours, whether they be intensive or dispersed over time. Instructional materials matter with an advantage

associated with the use of business simulation materials, and an even bigger edge associated with video materials. In-house training of the instructors' use of these materials serves to optimize the impact of the instruction.

The second conclusion we would like to underscore is that unanalyzed archival data may still be valuable for program evaluation. The key to retrieval of such data is identifying key characteristics of program organization that can be cast into survey protocols. These protocols can then be used to locate organizations with language training programs, with or without internal evaluation. Program characteristics can be reconstructed through the examination of curriculum documents and test records. Meta-analysis with ex post facto surveys can provide new interpretations of global program effects. Program designers can thus learn from retroactive analyses of previous program policies in evaluating options for future program improvement.

Finally, even though we have shown that archival data can be exploited to provide broad generalizations about organizational effects on language training programs, we still encourage program managers to conduct rigorous internal evaluation on an on-going basis.

Acknowledgements

We would like to thank The Institute for International Business Communications (IIBC) for access to the TOEIC Japan data sets, and The International Communication Foundation (ICF) for the TOEIC Korea data sets.

References

Alderson, C. and Berretta, A. (1992) *Evaluating Second Language Education.* London: Cambridge University Press.

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.

Hunter, J. and Schmidt, F (1990) *Methods of Meta-Analysis*. Newbury Park, CA: Sage.

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

Levin, H.M. (1983) *Cost-Effectiveness: a Primer.* Thousand Oaks, CA: Sage.

Lynch, B.(1996) *Language Program Evaluation* London: Cambridge University Press.

Ross, S. J. (2000) Individual differences and learning outcomes on the Certificate of Spoken and Written English. In G. Brindley (Ed.) *Studies in Immigrant English language Assessment*. Sydney: NCELTR.

Ross, S. J. (2003) A diachronic coherence model for language program evaluation. *Language Learning 53*, 1. 1-33.

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

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

Stuart, Alan (1991) *Kendall's Advanced Theory of Statistics.* New York: Oxford U. Press

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