TOEIC Research Report

NUMBER 1
SEPTEMBER 1989

Enhancing the Interpretation of a Norm-Referenced Second-Language Test Through Criterion Referencing: A Research Assessment of Experience in the TOEIC Testing Context

Kenneth M. Wilson

ETS
Educational Testing Service

Enhancing the Interpretation of a Norm-Referenced Second-Language Test Through Criterion Referencing: A Research Assessment of Experience in the TOEIC Testing Context

Kenneth M. Wilson

Educational Testing Service
Princeton, NJ

ETS RR-89-39

Abstract

This study was undertaken to develop guidelines for making interpretive inferences from scores on the Test of English for International Communication (TOEIC), a norm-referenced test of English-language listening comprehension (LC) and reading (R) skills, about level of ability to use English in face-to-face conversation, indexed by performance in the Language Proficiency Interview (LPI) situation. LPI performance, rated according to behaviorally defined levels on the LPI/ILR/FSI quasi-absolute proficiency scale, was treated as a context-independent criterion, using the familiar regression model in an apparently novel application (for such criterion-referencing purposes) in the context of a large-scale ESL-testing program. The study employed TOEIC/LPI data-sets generated during operational ESL assessments in representative TOEIC-use settings (places of work or work-related ESL training) in Japan, France, Mexico, and Saudi Arabia, involving samples of adult, educated ESL users/learners in or preparing for ESL-essential positions with companies engaged in international commerce. The pattern of TOEIC/LPI concurrent correlations was consistent across samples and there was relatively close fit between sample LPI means and estimates from TOEIC scores, especially TOEIC-LC, using combined-sample regression equations. Theoretical and pragmatic implications of the findings are discussed. General guidelines are provided for making inferences about LPI-assessed level of oral English proficiency from TOEIC scores. Directions are suggested for further research and development activities in the TOEIC testing context.

Summary

A generic problem with norm-referenced second-language proficiency tests is that examinees' scores on the tests do not provide a direct indication of their actual levels of functional ability to use a target language as demographically comparable native-speakers can be expected to use it. The functional implications of scores on such tests must be established empirically by conducting criterion-related validity studies designed to link level of performance on specific tests to level of performance on criterion measures of ability to use English (or other target language), operationally defined in some acceptable sense, based on direct observation and evaluation of the defined behavior.

The criterion observations may be either "context-specific" (e.g., samples of business correspondence, observation of communicative interaction with native speakers), or "context-independent" (e.g., ability to use English in face-to-face conversation, directly assessed using the Language Proficiency Interview [LPI] procedure that results in ratings of oral language proficiency according to inherently meaningful, behaviorally defined levels).

The Test of English for International Communication (TOEIC), developed by Educational Testing Service (ETS), is a multiple-choice, norm-referenced test designed to measure the English-language listening comprehension (LC), and reading (R) skills of individuals for whom English is a second language (ESL). The TOEIC is used primarily by corporate clients, worldwide; the majority of clients are located in Japan, as are about 80 percent of all TOEIC examinees. In Japan and in several other countries, TOEIC affairs are administered by local representative offices; elsewhere the TOEIC is available through the TOEIC-ETS (Princeton) office.

This study was undertaken to develop and evaluate guidelines for making inferences about level of oral English proficiency from TOEIC scores. Level of performance on the TOEIC was referenced to LPI ratings, for samples of examinees from representative test-use settings in Japan, France, Mexico, and Saudi Arabia, using the familiar regression model.

LPI ratings were regressed on TOEIC-LC, TOEIC-R, and TOEIC-Total, respectively, and on TOEIC-L and TOEIC-R (as a battery of predictors), in data-sets obtained under operational conditions. Study data were obtained during the course of comprehensive ESL proficiency assessments conducted by TOEIC-trained interviewers/raters in representative TOEIC-use settings in Japan, and by TOEIC/ETS staff members, including the staff member responsible for providing training in the LPI technique in Japan and elsewhere.

Across four Japanese subsamples (N = 42 through N = 142, combined N = 285), coefficients for Total/LPI ranged between .71 and .80, TOEIC-LC/LPI correlations ranged between .67 and .80; TOEIC-R/LPI coefficients were slightly lower (as expected on theoretical grounds), ranging between .65 and .72. Similar patterns of relationships were found in data-sets for samples of TOEIC examinees in France (N = 56), Mexico (N = 42), and Saudi Arabia (N = 10). The correlational findings indicated that inferences about LPI performance based solely on

TOEIC-LC were essentially as valid as inferences based on TOEIC-Total, the simple sum of LC and R, or on regression weighted composites of LC and R. Results of a residual analysis indicated that the fit between observed and estimated criterion means was more consistent across the four national samples, when LPI was estimated from a combined-sample regression equation using only the TOEIC-LC score, than from a combined-sample equation using the Total-score. These findings are evaluated from both theoretical and pragmatic perspectives.

Study findings suggest, as a strong working hypothesis, that level of ability to use English in face-to-face conversation (indexed by LPI performance) will vary relatively consistently with level of developed English-language listening comprehension (indexed by TOEICLC scores), across as well as within samples of educated, academically trained ESL users/learners likely to be tested with the TOEIC in diverse national TOEIC subpopulations. Guidelines for making interpretive inferences about levels of oral English proficiency from TOEIC scores are developed, and evaluated from theoretical and pragmatic perspectives. Attention is called to the problem of relating tested levels of proficiency to levels of on-the-job performance in positions that require the use of English, and to the problem of setting "minimum proficiency requirements."

It is concluded that by its initiative in encouraging and facilitating the use of the well-established LPI (direct assessment) procedure in operational testing contexts, the TOEIC program has made it possible to develop general guidelines that permit test users to make statistically valid inferences from TOEIC scores about levels of oral English proficiency. Furthermore, this initiative has made it possible to develop better-informed perspective regarding the level and range of developed oral English proficiency relative to expectation for an educated native speaker in the population of ESL users/ learners likely to be tested with the TOEIC. These are interpretive inferences that cannot be drawn from knowledge of distributions of standard scores on norm-referenced tests, alone.

Acknowledgments

The research reported in this paper was supported by the Test of English for International Communication (TOEIC) Testing Program office (ETS, Princeton). The TOEIC program office made this research possible by taking the initiative to facilitate the use of direct language proficiency assessment through the Language Proficiency Interview procedure, combined with its primary assessment instrument (the TOEIC) in settings where the TOEIC was already in use. The writer is deeply appreciative not only of the direct support provided by the TOEIC Program office at ETS, but also of the opportunity made possible by the availability of joint distributions of difficult-to-obtain LPI ratings and TOEIC scores for samples from representative TOEIC-use settings in four different countries to evaluate the utility of a logical, but largely neglected, model for setting functional guidelines for the interpretation of norm-referenced language proficiency measures in the context of a large-scale ESL proficiency testing program.

Special acknowledgment is made of the contribution to this undertaking, of Steven A. Stupak, director of the TOEIC Program office at ETS and the individual responsible for general oversight of TOEIC operations worldwide. Matthew Sindlinger of the TOEIC/ETS staff was consistently helpful in providing perspective regarding the nature of TOEIC operations in Japan and elsewhere; Gail Guadagnino coordinated communication with the TOEIC representative office in Japan during the course of the study.

Special thanks are extended to Akira Ito and his colleagues in the TOEIC Steering Committee in Japan: for providing TOEIC/LPI data-sets for samples of Japanese TOEIC examinees, for supplying essential information regarding the samples and the characteristics of the TOEIC examinee population in Japan, and for obvious meticulous attention to detail. Special acknowledgment is also made of the contribution of Vincent Reilley, Director of the English Language Program of the Japan-based Institute for International Studies and Training (IIST), and of his colleagues who generated LPI ratings for a substantial proportion of the basic Japanese calibration sample. Professor Reilley also provided clarifying detail about the practice of using the LPI procedure in combination with TOEIC scores in assessing outcomes of the IIST English Language Program.

John DeJong, Grant Henning, and Donald Powers provided helpful comments on a preliminary draft of the manuscript. Lawrence A. Stricker provided particularly helpful guidance in resolving several interpretive issues. Joanne Farr provided patient and typically competent assistance in the process of preparing and revising copy for the manuscript.

The foregoing contributions are gratefully acknowledged. However, the writer assumes full responsibility for the contents of this report; the views and interpretations expressed are his own, and are not necessarily shared by the TOEIC-ETS Program Office, by Educational Testing Service, nor by any of the individuals named above--no one other than the writer should be held responsible for deficiencies of expression, logic, or analysis, or for any other deficiencies that may be discerned in the manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

Section I: INTRODUCTION

A generic problem with norm-referenced tests of second-language proficiency is that the test scores do not provide any direct indication of actual levels of functional ability to use the target language(s) involved. As noted by Ingram (1985: 237), for example,

> [norm-referenced tests serve primarily] to discriminate amongst and rank-order learners and the learner's proficiency level is measured in relation to the performance of other learners, i.e., all one can directly say about the results of such tests is that on Test X Learner A was better or worse than Learner B or than n% of the other learners who took the test.

The interpretive problem has been succinctly summarized by Carroll (1967: 2), as follows:

> Except to the extent that one can guess at the range of competence possessed by a reference group, a percentile rank [on norm-referenced tests] cannot tell, for example, how successful the individual would be in communicating with a native speaker of the language or in comprehending the substance of printed materials in the language.

Thus, for example, knowledge of an examinee's standard scores or percentile ranks on a norm-referenced test such as the Test of English as a Foreign Language or TOEFL [ETS, 1985a], permits no direct inferences regarding the nonnative-speaker's functional ability to use English as a second language (ESL) or as a foreign language (EFL)--for example, to engage in a communicative dialogue with native-English speaking students or faculty members.[1] (see correspondingly numbered endnote here and hereafter).

Accordingly, norm-referenced tests of English language macroskills (e.g., listening or reading), or components of such skills (e.g., vocabulary), or knowledge of grammar, and so on, are referred to as <u>indirect</u> measures of "real-life language activities" (e.g., Clark, 1975: 10-11).

The functional implications of scores on such tests must be established empirically by conducting criterion-related validity studies designed to link level of performance on specific tests to level of performance on criterion measures of ability to use the target language, operationally defined in some acceptable sense, based on direct observation and evaluation of pertinent behavior. As noted by Clark (1975, 1978), for example:

> The usefulness (of indirect, norm-referenced tests) does not . . . depend on the tests' face/content validity but on the extent to which the test scores are found to correlate, on a statistical basis, with more direct measures of the proficiency in question (1978a: 27); (and) . . . the validity of indirect procedures as measures of real life proficiency is established through statistical—specifically correlational—means (1975: 11).

Although it seems clear that this is so, surprisingly little attention has been given to the exposition, evaluation, and application of models for conducting the types of criterion-related validity studies needed to establish the general level and consistency of concurrent relationships between scores on particular norm-referenced tests and specified criteria of ability to use English (or any other target language)--either general "context-independent" language-use criteria (for example, direct assessments of oral language proficiency in a controlled interview situation) or diverse "context-specific" criteria (reflecting observation and evaluation of ability to meet linguistic demands in various "real-life" work or study contexts). [2]

Thus, there is little direct precedent for the study reported herein--a study undertaken to establish interpretive guidelines for the Test of English for International Communication (TOEIC), by "calibrating" (a) scores on this norm-referenced ESL proficiency test to (b) behaviorally defined levels of "functional ability to use English in face-to-face conversation" (assessed formally in structured conversational interviews), treated as (c) a general, "context-independent" criterion variable, in (d) samples of TOEIC examinees from representative TOEIC-use settings in Japan and elsewhere.

For the present it is sufficient to establish the following points:

1. The TOEIC is a multiple-choice, norm-referenced test, with sections measuring English language listening comprehension and reading ability. The TOEIC testing program, developed and generally administered by Educational Testing Service (ETS), serves primarily corporate employers outside the United States who need to make English-proficiency-related personnel selection, placement, and/or training decisions (see, for example, ETS, 1982a, 1985b, 1986a, 1986b, 1988).

2. Functional ability to use English in face-to-face conversation was assessed using the well-established direct Language Proficiency Interview (LPI) procedure developed by the Foreign Service Institute (FSI) of the U.S. Department of State. Language Proficiency Interview is only one of several recognized designations for this direct oral language proficiency interview procedure, referred to originally as the Foreign Service Institute (FSI) Oral Proficiency Interview (OPI). The procedure has also been designated as the Interagency Language Roundtable (ILR) Oral Interview, reflecting the fact that it has been adopted by a number of U.S. governmental agencies, known collectively as the Interagency Language Roundtable (see Lowe, 1987). In the TOEIC testing context, the interview procedure is widely known as the LPI procedure. Regardless of the designation applied, this direct assessment procedure generates ratings of oral language proficiency according to inherently meaningful (behaviorally defined) levels on an "absolute proficiency scale" ranging from 0 (no proficiency) through 5 (proficiency equivalent to that of an educated native speaker [ENS] of the target language).

3. The familiar regression model was employed to calibrate (reference, link) scores on the arbitrarily defined TOEIC standard score scale to directly interpretable levels of LPI performance (treated as a "context-independent" language use criterion measure) in samples of TOEIC examinees in Japan and elsewhere.

Although it has been infrequently applied, the concept of setting general functional guidelines for the interpretation of norm-referenced second-language proficiency tests by calibrating the test scores to the directly interpretable LPI scale is logical, and it has strong empirical precedent. In a benchmark study of the attainments of foreign language majors in the United States, Carroll (1967) used a simple equating model to establish equivalencies between (a) scores on norm-referenced tests (of basic macroskills in French, German, Russian, and Spanish) and (b) LPI-scaled conversational interview ratings (and comparably scaled ratings of functional reading proficiency in the target languages), using data for (c) samples generally representative of the focal populations.

The elemental significance of the concept of enhancing the interpretation of norm-referenced language proficiency tests by referencing (calibrating) test scores to inherently meaningful, behaviorally-scaled direct proficiency measures--the basic concept embodied in Carroll's 1967 study design (the Carroll model)--apparently has not been generally recognized. In reviewing the research literature, for example, the writer was unable to find an extended discussion of the Carroll model, and no directly comparable study involving norm-referenced second-language proficiency tests appears to have been conducted in the United States.[3]

In circumstances such as those described, it is important to provide general context and perspective for the empirical study conducted in the TOEIC testing context along lines sketched above, by

1. considering briefly two complementary approaches to the design of criterion-related validity studies concerned with enhancing the interpretation of norm-referenced second-language tests, namely, studies involving "context-specific" (real-life) language use criteria, and studies involving general "context-independent" criteria (such as performance in Language Proficiency Interviews, the criterion employed in the study);

2. examining properties of the LPI that seem logically to establish the relevance of functionally scaled LPI behavior as a "context-independent" criterion for use in setting interpretive guidelines for indirect measures;

3. reviewing in some detail the pioneering study by Carroll (1967), and a large-scale study (Hilton, Grandy, Kline, & Liskin-Gasparro, 1985) in which self-assessments of oral language proficiency were calibrated to LPI ratings in samples of teachers of Spanish and French in the U.S.--studies that yielded important interpretive benefits by referencing scores on indirect measures (both test and nontest) to behaviorally scaled, direct measures of language proficiency, using simple equating models; and

4. detailing the advantages of a regression-based approach (over the equating approach) to calibrating the arbitrarily defined scales of norm-referenced, indirect proficiency measures to directly interpretable proficiency levels, using LPI-performance as a "context-independent" criterion variable.

A review of these elemental considerations is provided in Section II to establish the conceptual and methodological rationale for the empirical study in the TOEIC testing context that involved the use of a regression-based model for developing and evaluating the usefulness of guidelines for inferring (estimating) LPI performance from scores on the TOEIC, in samples of educated, adult ESL users/ learners from representative TOEIC-use settings (places of work or work-related intensive ESL training) in Japan, France, Mexico, and Saudi Arabia, using TOEIC/LPI data-sets generated during the course of comprehensive, operational on-site ESL assessments.

Study findings, in samples from the majority test-taking subpopulation in Japan, and in samples from three additional national test-taking subpopulations, indicate that clear interpretive benefits were realized by referencing scores on the TOEIC to LPI performance. The findings and other evidence reviewed in the study suggest, as a working hypothesis, that the pattern of TOEIC/LPI relationships observed in the study sample is likely to be relatively consistent across nationally and linguistically diverse samples of educated, adult ESL users/learners who are likely to be tested with the TOEIC.

Section II. ALTERNATIVE EMPIRICAL MODELS FOR ENHANCING THE INTERPRETATION OF NORM-REFERENCED TESTS

For purposes of the present paper it is useful to consider briefly the principal characteristics of two complementary types of criterion-related validity studies, namely, studies designed to relate scores on norm-referenced (indirect) tests to "context-specific" criteria of ability to use a target language, and those designed to relate the test scores to "context-independent" criteria.

Relating Indirect Tests to Context-Specific Criteria

In studies involving context-specific criteria, the aim is to relate scores on norm-referenced tests to criteria that reflect functional ability to use a target language in specific settings (e.g., places of work or study) to perform language-essential tasks (e.g., discuss business affairs with native speakers or participate in an academic seminar; write business letters or term papers). Performance might be assessed by native-speaking supervisors, colleagues, or clients. Studies of this type have been characterized as "ultimate pragmatic validity studies" (Ingram, 1985: 238).

By linking score levels on the tests to context-specific criteria, local users can obtain the type of evidence that is needed to help them form realistic (actuarially based) expectations about the type or level of on-the-job language proficiency likely to be exhibited by individuals at different score levels on the test under consideration. The context-specific approach provides interpretive guidelines that are locally meaningful.

However, factors that make context-specific studies valuable for local test users tend to militate against generalization. For example, many replications of studies involving particular tests and criteria would be needed to assess the stability of relationships across contexts. Moreover, context-specific criteria--like the indirect test being "pragmatically validated"-- are likely to involve only a relativistic classification of the linguistic behavior being evaluated (e.g., superior, average, below average; satisfactory versus unsatisfactory). Thus, despite their local pragmatic value, the results would not contribute directly to improved understanding of the general levels or types of "ability to use a target language" that individuals at specified score levels on the norm-referenced test may be expected to exhibit.[4]

Finally, it is difficult for professionals to design and conduct rigorous follow-up studies; local test-users are likely to find it even more difficult (conceptually and logistically) to do so. Consequently, as Ingram (1985: 238) has noted, " . . . few if any adequate studies exist relating indirect tests to real life or workplace use of the language."[5]

Generally speaking, in pragmatic validation involving context specific criteria, the questions at issue have to do with whether individuals at given score levels on a test are linguistically qualified for particular ESL-essential jobs, with how adequately they perform the ESL aspects of their work, with identifying minimally acceptable standards, and so on.

Relating Indirect Tests to Context-Independent Criteria

In order to obtain more general answers to questions about the functional implications of scores on particular indirect proficiency tests, it is necessary to conceptualize and conduct studies employing "context-independent" criteria.

A context-independent criterion may be defined as a measure of ability to use the target language in circumstances resembling those likely to be routinely encountered in many different "real-life" language-use contexts (e.g., situations requiring the exchange of meaning in conversational interaction). In context-independent studies, the aim is to assess the relationship between scores on particular norm-referenced tests and level of performance on one or more criteria of "functional ability" to use a target language, rated according to a scale involving behaviorally defined levels.

Given scores on the clearly defined, functionally scaled criterion variable and scores on an indirect, norm-referenced test for a representative sample from a defined population of second-language users, application of the familiar regression model would make it possible to translate (calibrate) scores on the arbitrarily defined scale of the norm-referenced test into estimated scores on the behaviorally scaled, hence directly interpretable, criterion through a regression (calibrating) equation. Questions regarding the applicability of the resulting regression equation across subpopulations of interest (e.g., different native-language subgroups) can be addressed empirically (e.g., through residual analyses designed to assess average discrepancies between observed standing on the functional criterion and estimated standing based on the general regression equation).[6]

Behavior in Language Proficiency Interviews as a Context-Independent Criterion

The LPI model was developed for use in assessing the linguistic readiness of personnel to undertake assignments with the U. S. government in posts requiring particular levels of oral language proficiency in languages other than English. The levels specified by the model are used to characterize, in functional terms, both the individual's performance and the demands of particular positions. The LPI procedure has been adopted without basic modification by nongovernmental institutions and agencies in both public and private settings for purposes of evaluation (of second-language training) and certification (e.g., of second-language teachers).[7]

The behavior that is elicited under controlled conditions and systematically rated in the Language Proficiency Interview appears to be similar to the type of behavior that is elicited (and evaluated or judged informally) in a variety of real-life contexts. The following description of the procedures is provided by Clark and Swinton (1979):

> "The interview consists of a face-to-face conversation of approximately 15-25 minutes between the examinee and a trained interviewer who is a native or near-native speaker of the test language. The conversation begins at a fairly simple level and becomes increasingly more sophisticated linguistically, as reflected in increased rate of speech on

the part of the interviewer, use of more complex structures and more specialized vocabulary, up until the point at which the examinee is no longer able to hold his or her own in conversation at that level. At this point, the level of sophistication of the conversation is reduced somewhat and the examiner spends several minutes exploring the examinee's breadth of command of grammatical structures (for example, ability to use past and future tense forms, conditional constructions, etc.); and extent of active vocabulary, as elicited by questions probing a variety of topical areas including personal and family background, work activities, studies, hobbies and free-time activities, future plans, and so forth. With more proficient examinees, the interviewer will also broach political, social, economic, or other topics requiring very high levels of language use. The interview continues until the examiner is satisfied that the interviewee has fully demonstrated the highest level of speaking proficiency of which he or she is capable" (p. 5).

The FSI/LPI "Absolute Proficiency Scale"

The most distinctive feature of the LPI model is the scaling frame of reference employed. The behavior assessed is classified by trained interviewers/raters according to levels ranging from "0" (indicating no functional language-use ability in the `situation) through "5" (indicating functioning equivalent to that of an educated native speaker [ENS]).

Each of six principal points (0,1,2,3,4,5) on this "quasi-absolute proficiency scale" (to use Carroll's [1967: 2] apt modification of the typically employed term "absolute") is "anchored" behaviorally. Each level is characterized by a clearly defined pattern of language-use behavior. In traditional score-reporting practice, for levels 0-4 a "+" is added to a level-rating for individuals whose performance is judged to substantially exceed that for a given level, but not to meet fully the requirements for the next higher level. In analyses requiring numerical conversions, the plus ratings are designated by adding .5 to the level whose requirements have been met fully (O+ = .5, 1+ = 1.5, and so on).[8] Detailed descriptions of the type of linguistic behavior associated with each of the LPI levels (basic and intermediate) is provided in Appendix A.

Interpretive inferences from rated levels. The types of interpretive inferences associated with each of the basic levels of the LPI scale have been succinctly summarized in the form of a "functional trisection," (shown as Exhibit A), conceptualized by an agency of the U.S. government (see ETS, 1982: 21). The trisection characterizes each level as to type of functional ability demonstrated, content areas covered, and level of accuracy of language use in the interview situation. Level 3 has come to be accepted as indicating "minimum professional mastery" of a second language. From a normative perspective, according to one informed observer of the LPI procedure (Jones, 1978: 93), circa 1978 (a) there were very few examinees above Level 3, and (b) very few language-essential positions within the U.S. government that were designated as requiring proficiency higher than Level 3.[9]

LPI Performance as a General Context-Independent Criterion

It has been suggested by second-language assessment experts (e.g., Clark, 1975) that face-to-face conversation approaches real-life communication about as closely as is possible in a test

Exhibit A

FUNCTIONAL TRISECTION OF ORAL PROFICIENCY LEVELS*

Oral Proficiency Level	Function (Tasks accomplished, attitudes expressed tone conveyed)	Context (Topics, subjects areas, activities and jobs addressed)	Accuracy (Acceptability, quality and accuracy of message conveyed)
5 (Superior)	Function equivalent to an educated native speaker (ENS).	All subjects .	Performance equivalent to an ENS.
4 (Superior)	Able to tailor language to fit audience, counsel, persuade, negotiate, represent a point of view and interpret	All topics normally pertinent to professional needs.	Nearly equivalent to a ENS. Speech is precise, appropriate to every occasion with only occasional errors.
3 (Superior)	Can converse in formal and informal situations, resolve problem situations, deal with unfamiliar topics, provide examinations, describe in detail, offer supported opinions, and hypothesize.	Practical, social professional and abstract topics, particular interests, and special fields of competence.	Errors never interfere with understanding and rarely disturb the native speaker. Only sporadic errors in basic structures.
2 (Advanced)	Able to fully participate in casual conversations, can express facts, give instructions, describe, report, and provide narration about current, past and future activities activities.	Concrete topics such as own background, family, interests, work, travel, and current events.	Understandable to native speaker not used to deal ing with foreigners. Some times miscommunicates.
1 (Intermediate)	Can create with the language, ask and answer questions, participate in short conversations.	Everyday survival topics and courtesy requirements.	Intelligible to native speaker used to dealing with foreigners.
0	No functional ability	None.	Unintelligible.

*From ETS (1982a); see also Thompson (1985: 151)

situation. The relevance and utility of formally elicited and evaluated interview behavior as a "surrogate" for direct observation and evaluation of ability to exchange meaning conversationally in situations that arise naturally in a variety of workplace, academic, or other language-use contexts appears to be inferable directly from the procedures described above. Moreover, substituting the controlled interview for observation and assessment of the ". . . operational (language-use) capability of a man on the job" was proposed by Francis Cartier (1975), a discussant of Wilds' (1975) frequently cited seminal paper describing the development and use of the "oral interview test." Noting problems involved in obtaining real-life criteria, Cartier commented as follows:

> Let me point out that without at least metric access to the criterion situation, we have [in the structured interview] what we must call a surrogate criterion. We would like to, for example, correlate paper and pencil tests with interviews, and the reason we would like to do that is that the interviewer gives us this kind of surrogate criterion which we have to use simply because we can't apply any sort of metric to the criterion population and situation (Cartier, 1975: 12).

It is perhaps obvious that the interview situation does not provide a basis for simulating language-use requirements in all possible "real life" contexts. As noted by Clark and Swinton (1979: 6):

An obvious shortcoming is that the interview setting cannot directly reproduce the [great variety of] physical surroundings in which the examinee would be expected to perform in real life. . . (T)he psychological and affective aspects of real-life communication, including motivation and communicative intent of the speakers, status roles of interviewer and interviewee, and a number of other aspects of the real-life situation cannot be precisely duplicated in the interview setting (Clark & Swinton, 1979: 6)

Limitations of this kind may be thought of simply as inherent constraints in generalizing "level of second-language conversational ability" from the structured interview results--results that should not be expected to be equally predictive of functioning (language-use criteria) in every real-life situation involving the exchange of meaning through conversation. After all, on the basis of Cartier's succinct conceptualization of the issue, we may say that the controlled interview technique is useful as a surrogate criterion for referencing scores on indirect tests precisely because it is not possible to measure language-use behavior generally conceived.

Acceptance of LPI performance as a surrogate for "real-life" performance criteria for validating "paper and pencil" (multiple-choice, norm-referenced, indirect) tests does <u>not</u> obviate the need for pragmatic validation of the surrogate criterion itself, using some measure of "operational capability of a man on the job" (especially of the language-use dimensions of such capability) as a criterion.[10]

Reliability Considerations

The reliability of the LPI-criterion was not directly at issue in this criterion-referencing study.

"High reliability in a criterion measure is convenient but not critically important. Low reliability in a criterion measure merely attenuates all its relationships with other measures" (Thorndike, 1949: p. 127).

However, there is a significant body of empirical evidence bearing on the reliability of the LPI procedure as administered within the U.S. government (e.g., Adams, 1978; Clark, 1978a, 1978b passim) and elsewhere (e.g., Clark, 1978c; Clark and Swinton, 1979; Hilton et al. 1985).

For purposes of the present study, it is useful to call attention to certain general conditions that have been found to affect the "reproducibility of LPI ratings"--that is, consistency with regard to both rank-order and level in rating LPI performance. The number of raters, of course, is a generic reliability-related factors--reliability tends to increase as the number of raters increases.

Apart from this generic consideration, the reproducibility of ratings, by raters trained in the LPI procedure, is enhanced when all raters "share the same roof." As noted by Adams (1978: 35), the system works best with all interviewers ". . . under one roof, able to consult with each other . . . and most apt to break down . . . when examiners are isolated." Limited empirical evidence of the effects of "same site" versus "scattered site" conditions on the reproducibility of LPI ratings tends to support Adams' observation.

Clark (1987) compared ratings (in French and German, respectively) for the same interviewees by interviewer/raters in three different U.S. government agencies.

> "(Although) the ratings assigned did not differ across agencies in a statistically significant way . . . examination of the rating performance for various sub-portions of the proficiency scale showed fairly clear across-agency differences . . . primarily at the lower and middle ranges of the scale" (p. 145). (Clark did not examine intra-agency reliability.)

Bejar (1985) found that reliability of ratings of samples of ESL-speaking behavior, represented by taped recordings of items from the Test of Spoken English (ETS, 1985c), improved when "same site" conditions were introduced.

The Carroll Model

Like any assessment procedure that involves direct observation of individual behavior, and clinical or subjective evaluation of observed behavior samples, the LPI model is too costly and cumbersome to administer to be considered as the primary instrument in large-scale programs for which the multiple-choice, indirect, norm-refer-referenced test is admirably suited. However, as indicated earlier, Carroll (1967) recognized that the interpretive power of behaviorally anchored, direct assessment procedures such as the LPI (and parallel procedures for assessing reading or writing skills) could be extended to populations of interest by empirical linkage to related, norm-referenced tests, using linkage rules developed in samples from the populations.

No other large-scale studies using the basic Carroll (1967) model to calibrate norm-

referenced test scores to the LPI scale appear to have been conducted in the United States. However, the Carroll model was employed to calibrate self-ratings of oral language proficiency to the LPI scale in a national study of the oral language proficiency of teachers of French and Spanish in the U.S. (Hilton, Grandy, Kline, & Liskin-Gasparro, 1985). For purposes of the present study, therefore, it is quite important to examine Carroll's conceptual and methodological approach, as well as illustrative findings from both of these national studies, in some detail.

Caroll's (1967) "Calibration" Study

Carroll was interested in assessing the foreign language skills of language majors near the end of their senior year in college and, incidentally, in developing national norms for a series of proficiency tests known as the Modern Language Association Proficiency Tests (MLAPT) in French, German, Italian, Russian, and Spanish. Each norm-referenced test included measures of listening comprehension, speaking (scored by trained judges), reading, and writing (a free response "cloze" type of test, scored by trained judges).

To address the problem of inferring language-use ability from scores on the norm-referenced MLAPT, a "calibration sub study" was conducted. The purpose of this sub study was

> ". . . to ascertain correspondences between MLA Proficiency Test scores and the 'absolute proficiency ratings' rendered by expert teams from the Foreign Service Institute of the U.S. Department of State--(that is) to calibrate the scores on the MLA Proficiency Tests in terms of `quasi-absolute,' inherently meaningful standards" (Carroll, 1967: 2).

To establish the correspondence between the test scores and ratings (LPI or Speaking [S] and Reading [R]), Carroll initially hoped to obtain data for samples of about 50 cases in each language. Ratings were finally obtained for somewhat smaller samples composed of participants in summer language institutes (attended by teachers and advanced students). The basic data generated in the equivalency sub study are summarized in Exhibit B: Carroll's (1967) Table 2.2, showing n's, means, standard deviations, and intercorrelations of MLAPT scores and ratings for the respective calibration samples (reprinted by permission of the Harvard Graduate School of Education). A brief description of the levels for FSI Speaking (that is, FSI/LPI levels) and the corresponding levels for FSI Reading (R-scale) is included in the table.

Carroll made a decision to link MLAPT speaking and listening scales to the Speaking scale (S-scale), and the MLA reading and writing scales to the Reading scale (R-scale). The variables were relatively highly intercorrelated.[11] Regarding this decision, Carroll made the following observation:

> "The correlations between the two FSI ratings, S and R, are quite high . . . Save possibly in the case of French, there is little evidence in the FSI-MLA correlations to suggest that FSI Speaking [LPI] ratings are more highly correlated with MLA Listening and Speaking scores than with MLA Reading and Writing scores, nor that FSI Reading ratings are more correlated with MLA Reading and Writing scores than they are with Listening and Speaking scores. Nevertheless, on an a priori basis [the

linkage pattern designated above was followed]" (Carroll, 1967: 13).

On the basis of the data summarized in Exhibit B, MLAPT scores were translated to either the LPI/S-scale or the corresponding scale for reading (the R-scale, for which general descriptors are provided in Exhibit B), in the pattern indicated above, by the method of "equal standard scores": *{z(x) = [X X*] / SDx} = {z(y) = [Y Y*]/ SDy},* where **X** and **Y** are observed scores, **X*** and **Y*** are means, and **SDx** and **SDy** are standard deviations of the respective variables.

Carroll commented on his use of an equating model for calibrating the MLAPT scores to the functional scales of the criterion variables, as follows:

> "(The equating approach) . . . merely assumes that X and Y are equally estimates of the same thing, and that it is an arbitrary matter whether one measures this thing by X or by Y . . . The more X and Y are correlated, the more this procedure is justified. It is felt that in the present case, the corresponding measurements are sufficiently well correlated to justify the procedure, particularly in view of the fact that the purpose of the study was merely to establish meaningful standards for the interpretation of MLAPT scores" (Carroll, 1967: 15, emphasis added).[12]
>
> The observed means and standard deviations of the ratings in the calibration samples permitted (probably for the first time) empirically based inferences about the level of development of significant aspects of functional ability to use particular target languages in particular populations of second language learners (relative to expectation for educated native-speakers). By calibrating MLAPT scores to the inherently interpretable scales of the direct assessments, Carroll and his associates were able to extend these inferences to the populations of interest.

For purposes of the present study it is useful to examine selected patterns of findings that point up the interpretive contribution of referencing scores on the norm-referenced MLAPT-series to the functionally defined LPI (Speaking) scale and the conceptually comparable Reading scale.

Interpretive contribution--some illustrative examples. In evaluating results for college-senior-level majors in the respective languages, Carroll (1967: 46), commented as follows:

> Taking the results at their face value . . . we find that a general characteristic of the tested samples is that they are much poorer in Listening and Speaking skills than they are in Reading and Writing.

This pattern is evident in Figure 1a (showing the pattern of functionally scaled equivalents of the MLAPT medians for the four groups of majors), and Figure 1b (showing relative frequency distributions for estimated functional Speaking (LPI) and Reading levels for French majors only.[13] The level indicating "minimum professional proficiency" (Level 3) is highlighted.

Exhibit B
Data for the Calibration Samples: Table 2.2 from Carroll (1967: 14)

	FRENCH		GERMAN		RUSSIAN		SPANISH	
No. tested	39		39		19		30*	
Test	Mean	S.D.	Mean	S.D.	Mean	S. D.	Mean	S.D.
MLA List.	47.38	6.07	45.62	7.93	41.84	5.48	44.57	5.71
" Speak.	82.97	9.84	97.90	19.83	85.00	10.92	84.87	9.49
', Read.	50.41	7.82	51.59	11.03	33.16	8.96	43.77	5.94
" Write	49.05	8.10	55.51	14.26	56.32	9.31	52.83	10.72
FSI Speak.*	2.62	.64	3.13	1.08	1.97	.66	2.58	.75
FSI Read.*	3.15	.66	3.10	1.10	1.89	.57	2.86	.66
Correlations with FSI Ratings								
	"S"	"R"	"S"	"R"	"S"	“R”	"S"	“R”
MLA List.	.67	(.61)	.73	(.72)	.84	(.75)	.73	(.80)
" Speak.	.67	(.49)	.82	(.83)	.78	(.66)	.66	(.65)
" Read.	(.58)	.71	(.82)	.82	(.78)	.69	(.63)	.74
" Write	(.65)	.63	(.86)	.84	(.62)	.71	(.70)	.77
FSI "R"	(.69)		(.95)		(.90)		(.80)	

* In computing these values, a "+" is given a value of .5. Thus, 1+ is coded 1.5, 2+ - 2.5, etc. For the meanings of the FSI ratings, see below.

S-5 Speaking proficiency equivalent to that of an educated native speaker.
S-4 Able to use the language fluently and accurately on all levels normally pertinent to professional needs.
S-3 Able to speak the language with sufficient structural accuracy and vocabulary to satisfy representation requirements and handle Professional discussions within a special field.
S-2 Able to satisfy routine social demands and limited office requirements.
S-1 Able to satisfy routine travel needs and minimum courtesy requirements.
R-5 Reading proficiency equivalent to that of an educated native speaker.
R-4 Able to read all styles and forms of the language pertinent to professional needs.
R-3 Able to read non-technical news items or technical writing in a special field.
R-2 Able to read intermediate lesson material or simple colloquial texts.
R-1 Able to read elementary lesson material or common public signs.

"All the ratings except the S-5 and R-5 may be modified by a plus (+), indicating that proficiency substantially exceeds the minimum requirements for the level involved but falls short of those for the next higher level."

--Extracted from "Absolute Language Proficiency Ratings," Circular, May 1963, Foreign Service Institute, Washington, D.C.

Figure 1a. Estimated functional proficiency levels corresponding to MLA medians for U.S. college seniors majoring in French,German, Russian, or Spanish: Adapted from Carroll (1967)

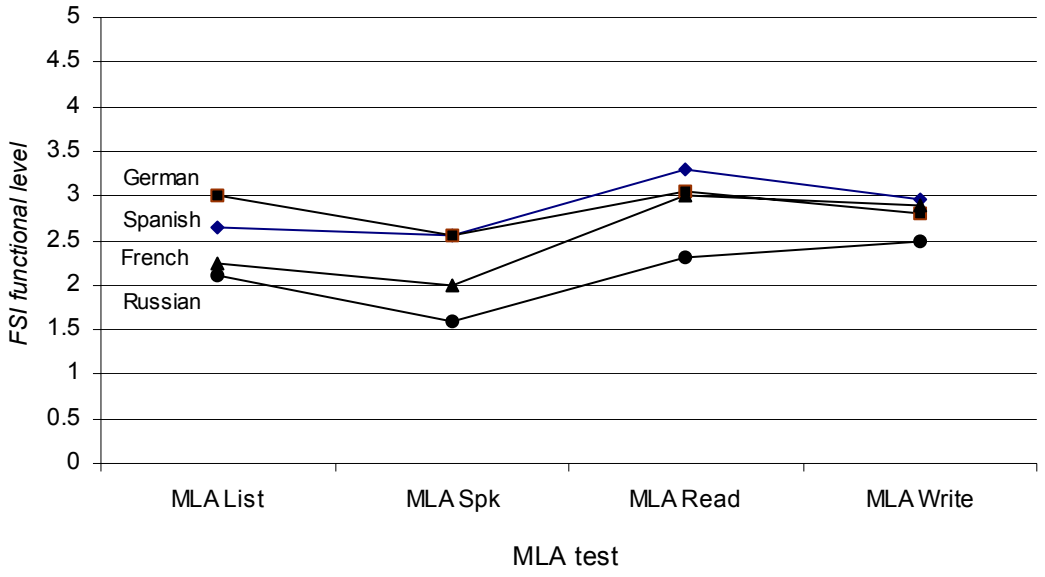
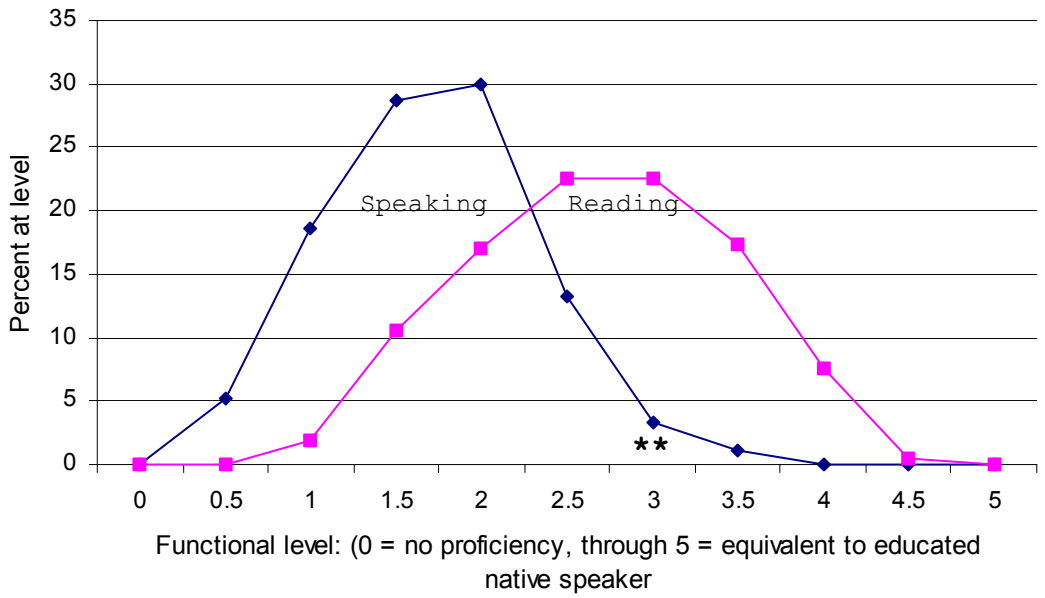


Figure 1b. Estimated distribution of Speaking and Reading ratings for French majors (adapted from Cqrroll, 1967)



The findings reflected in these figures permit inferences about the differential development of functional second-language listening and reading skills in the populations of interest. It is important to recognize that inferences about differential levels of functioning with respect to specified language skills in a given population of nonnative speakers cannot be derived from consideration of average standard scores on norm-referenced tests of listening and reading skills in representative samples of users/ learners from the intended population of test takers. In order to make such inferences, it is necessary to contrast the performance of samples from the focal (test-standardization) population on the respective measures with that of native speakers of the target language.

This point is reinforced in Figures 2a, 2b, & 2c. The figures are based on score-distributions reported by Angoff and Sharon (1971), who studied the comparative performance on the Test of English as a Foreign Language (TOEFL) of a sample of U.S. college freshmen in a relatively unselective college with that of a general TOEFL reference group. TOEFL examinees typically have relatively high levels of academic/cognitive skills developed primarily through the medium of their respective native languages.

Figure 2a shows distributions of standard scores for TOEFL Listening Comprehension and TOEFL Reading Comprehension for a general sample of TOEFL examinees (who took the five-section edition of the test). The two distributions are identical for all practical purposes. This is a psychometrically assured phenomenon. Because both sections were standardized in a sample from the population of interest, the listening comprehension items and the reading items were specifically selected so as to be of "average difficulty" for the standardization sample. This process effectively obscures any developmental differences that may be present in a target population.

When the performance of a group of native-English speakers on the respective sections is introduced as an interpretive frame of reference, differential levels of skill development are clearly inferable (cf. Figures 2b and 2c).

In essence, the general nature of the inferences from the series of figures is similar to the general nature of inferences from Figure 1b, reflecting Carroll's directly interpretable findings. It is not possible to draw comparable functional inferences from the distributions of TOEFL Listening and Reading scores for a general TOEFL reference group, as indicated in Figure 2a.[14] This is a generic problem with norm-referenced tests of second-language proficiency.

Calibrating Self-Assessments to the LPI Scale

Hilton, Grandy, Kline, & Liskin-Gasparro (1975), with the collaboration of Steven A. Stupak and Protase E. Woodford (ETS staff members), conducted a study of the oral language proficiency of foreign-language teachers in the U.S., in which an equating approach similar to that employed by Carroll (1967) was used to reference self-ratings of oral language proficiency to the LPI scale.

Figure 2a. Performance of nonnative-English speakers on TOEFL Listening Comprehension and TOEFL Reading

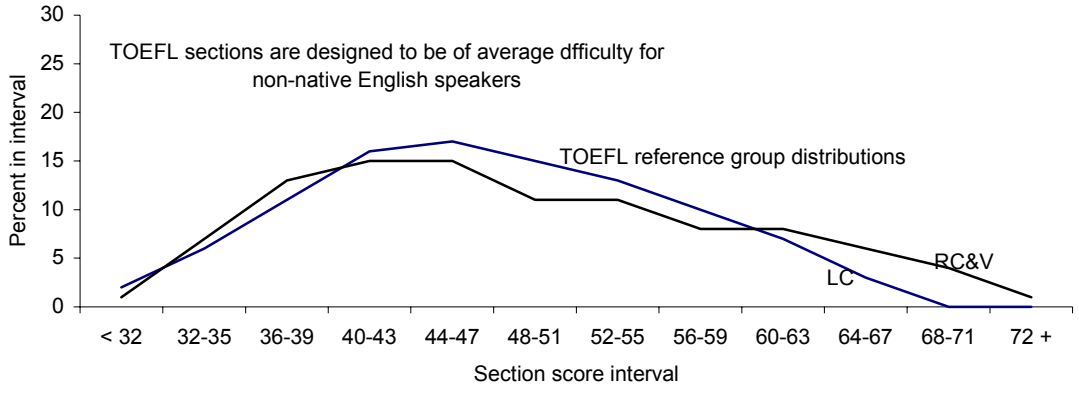
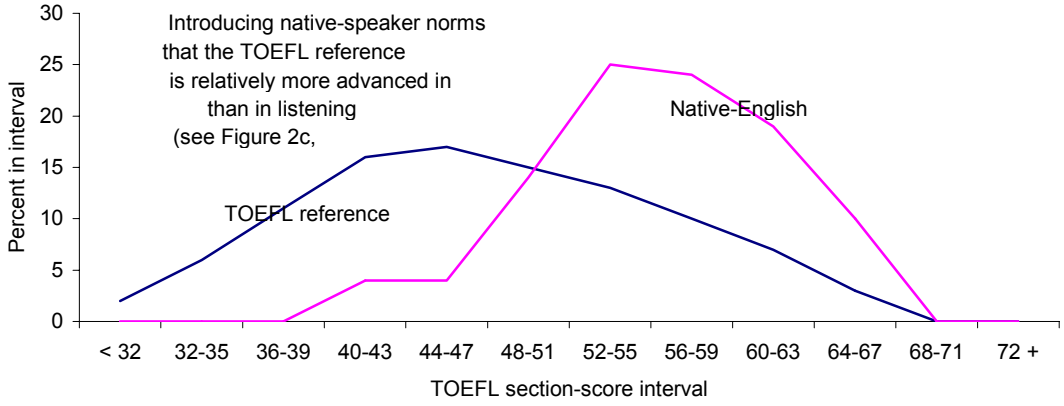
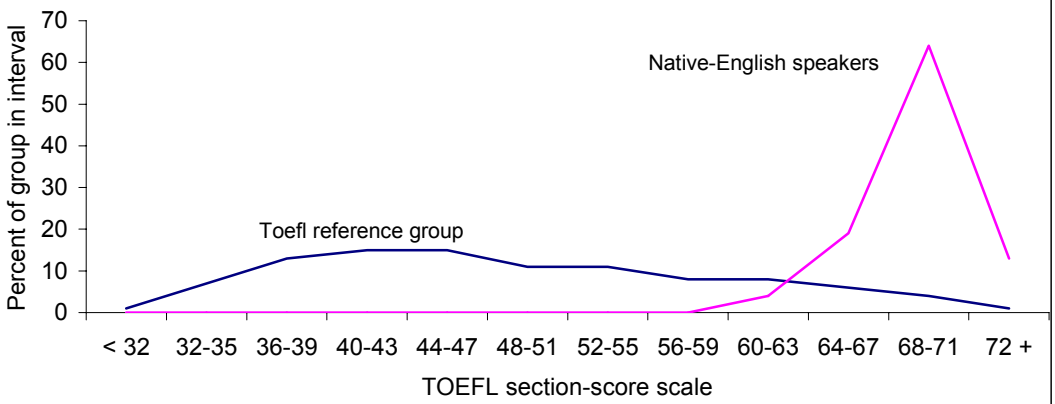


Figure 2b. Comparative performance of native- and nonnative-English speakers on TOEFL Reading Comprehension



Err

Figure 2c. Comparative performance of native- and nonnative-English speakers on TOEFL Listening Comprehension



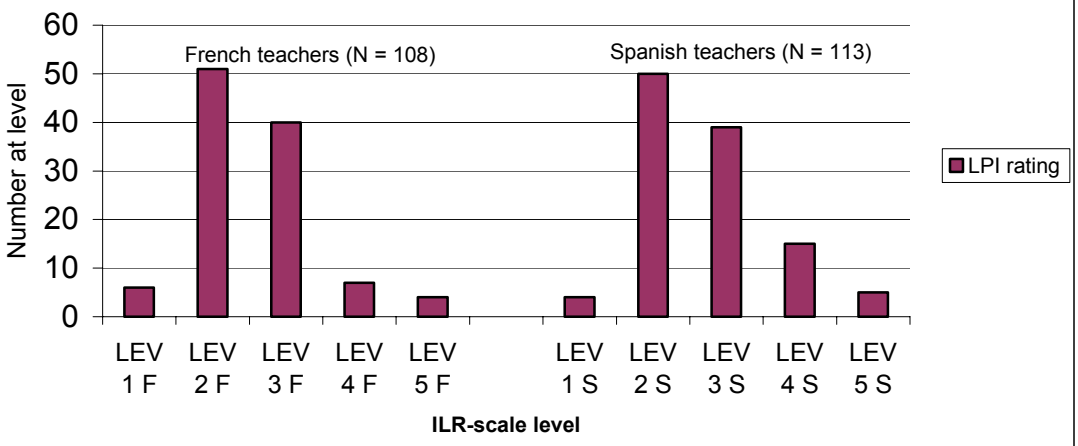
The Calibration Substudy

LPI ratings were obtained for 108 teachers of French and 113 teachers of Spanish, using recorded interviews. A total of 27 field raters and eight ETS language-staff members were involved in the ratings; at least two ratings were obtained for each interview.

Average inter-rater reliabilities for the two "regular" raters were .71 (French) and .73 (Spanish)—a "master" rater was used when regular raters disagreed in level by more than one-half point.[15]

Figure 3 shows relative frequency distributions of LPI ratings for the two validation samples. The distributions appear to be somewhat positively skewed and centered below the "minimum professional proficiency" level (LPI Level 3). This appears to represent a relatively low level of functional ability in using target languages in populations that may reasonably be assumed to be highly selected in terms of developed functional proficiency.[3]

Figure 3. Distributions of LPI ratings for teachers of French and Spanish, respectively (adapted from Hilton et al., 1985)



Note. Excerpts from Functional Trisection (see Exhibit A, above).

Level 5 Functions equivalent to an educated native speaker.

Level 4 Able to tailor language to fit audience . . . on all topics pertinent to professional needs

Level 3 Can converse in formal and informal situations . . .deal with unfamiliar topics . . . offer supported opinions.

Level 2 Able to participate fully in casual conversations, can speak in extended discourse, express facts, give instructions, describe, report, and provide narration about current, past and future activities.

Level 1 Can create with the language, ask and answer questions, participate in short conversations.

Level 0 No proficiency

The results of an analysis of potential correlates of these criterion ratings suggest that many (if not the great majority) of the higher rated members of the validation ("calibration") samples were probably native speakers of the target languages, not native-English speakers academically trained to pursue a career in teaching French or Spanish.

Of the more than 50 background variables included in the study, those listed below had the highest correlations with the oral language proficiency criterion (results are for the combined sample of French and Spanish teachers):

Background variables	Correlation with rating
Parents' native language (target language)	.57
Birthplace (country of target language)	.53
Native language (target language)	.52
Time spent in country where target language is the dominant language	.41
Speak target language at home	.33
Speak target language with spouse	.30
Spouse's native language (target language)	.29

The correlations cited, above, suggest that the highly rated teachers probably were native speakers of the languages under consideration. In sharp contrast, training-related variables were not very predictive of proficiency ratings: for example, college French grades (-.05), years of language beyond high school (.09), years teaching foreign language (.07), opportunity to continue study of present language (.07), and so on.

Global self-ratings were calibrated to the LPI scale by the method of equipercentile equating (Hilton et al., 1985: 26-27), with the implicit assumption, expressed explictly by Carroll, of functional equivalence for study purposes. Conclusions regarding the distributions of oral language proficiency in the general samples of teachers were similar to those for the calibration sample.

Contribution of Studies Using the Carroll Model

The findings of Carroll's benchmark study, reinforced by those of Hilton et al. using the Carroll model, are elementally important because they point up the intrinsic interpretive value of the models developed by the Foreign Service Institute for the direct assessment of oral English proficiency (the LPI) and reading proficiency, and their corresponding "quasi-absolute" proficiency scales, when applied to samples from defined populations of second-language users/learners. The findings also indicate clearly that the interpretive value of indirect measures can be established empirically by recalibrating the scales of norm-referenced tests (or other indirect measures, including self-ratings of oral language proficiency) to the behaviorally anchored LPI scale (or the conceptually comparable ILR/FSI Reading scale).

Knowledge of the distribution of LPI ratings for samples from a population of users/learners provides meaningful information regarding the probable level and dispersion of a clearly defined functional ability to use the language under consideration in that population.

Such inferences cannot be drawn from knowledge of sample distributions of scores on norm-referenced proficiency measures, alone.

Carroll's (1967) pioneering study demonstrated clearly that knowledge of the joint distributions of ratings based on direct assessment procedures and norm-referenced test scores in samples generally representative of a focal population can be used to establish "linkage rules" that provide a basis for inferences about level of functional ability to use a target language, from examinees' scores on norm-referenced tests--the principal interpretive issue. Hilton et al. (1985) demonstrated the generalizability of the Carroll model by calibrating self-assessments of oral language proficiency to the LPI scale.

In both of these large-scale empirical studies, the linkage rules employed involved a working assumption that the indirect and the direct measures were functionally interchangeable for the purpose of providing interpretive guidelines for indirect measures. To reiterate Carroll's (1967) characterization, use of an equating model " . . . merely assumes that X and Y are equally estimates of the same thing and that it is an arbitrary matter whether one measures this thing by X or by Y. The more X and Y are correlated, the more this procedure is justified."

As shown earlier (in Exhibit B), for Carroll's four relatively small calibration samples the correlations of the MLAPT Listening and Speaking scores with LPI (Speaking) ratings (across language groups--French, German, Spanish, Russian) ranged between .66 and .84. Comparable ranges for the MLAPT Reading and Writing scores versus Reading ratings were, respectively, .69 to .82, and .63 to .86. Correlations between Speaking and Reading ratings were .69 (French), .95 (German), .90 (Russian), and .80 (Spanish). And, of course, Carroll concluded that for the purpose of setting interpretive guidelines for the MLAPT, levels of intercorrelation such as the foregoing were quite satisfactory.

Advantages of the Regression Model

Despite the demonstrably improved interpretive perspective provided by equating models, such as those employed by Carroll (1967) and Hilton, et al. (1985), it is preferable to employ an approach to linking performance on indirect, norm-referenced tests to levels of performance on FSI-scaled language-use criteria that does not require the assumption of equivalency, even for working purposes.

Given joint distributions of LPI ratings and scores on indirect, norm-referenced measures for a given sample, it is clear that a regression-based calibration model does not require a priori assumptions about the organization of second-language skills, or the psychometric or theoretical equivalence of the measures involved.

At the same time, a regression-based approach to this problem obviously need not be atheoretical. By regressing LPI ratings on measures of listening and reading skills, for example, it is possible to assess the hypothesis of greater correspondence between second-language

speaking and listening skills than between speaking and reading skills, while at the same time establishing and evaluating statistically meaningful criterion-estimation rules.

In this connection, it is noteworthy that in the regression model, but not in the equating model, the scales of the indirect measures involved are referenced (calibrated) to the functionally scaled criterion variable according to linkage rules that vary directly with the observed level of association between the indirect measures and the functional criterion in calibration samples. Thus, regression-based estimates of criterion behavior are more explicitly "delimited" than are inferences that derive from the application of simple equating models. And, the usefulness of the regression model for purposes of criterion-referencing is well established.

As a general proposition, regressing a functionally scaled criterion variable of the type represented by LPI performance, on indirect, norm-referenced test scores in samples of test takers from defined populations of second-language user/learners, can be expected, a priori, to provide evidence that permits an informed evaluation of the patterns of relationships among the measures under consideration from both theoretical and practical perspectives, statistically delimited inferences (e.g., estimates, with standard errors), from scores on the indirect test, about probable level of defined language-use behavior, for individuals in samples from the test-taking population involved, and inferences regarding the probable level and dispersion of oral language proficiency in the test-taking population, according to the directly interpretable LPI scale.

Section III. REFERENCING TOEIC SCORES TO LPI RATINGS IN THE TOEIC TESTING CONTEXT

The empirical study described in this section was undertaken to provide an assessment of efforts by the TOEIC testing program to establish functional guidelines for interpreting scores on the Test of English for International Communication (TOEIC), a norm-referenced test of English-language listening comprehension (TOEIC-LC) and reading skills (TOEIC-R), by relating TOEIC scores to levels of "functional ability to use English in face-to-face conversation," defined operationally by behavior elicited using the Language Proficiency Interview (LPI) procedure, and rated according to the behaviorally anchored LPI (oral language proficiency) scale.

The study involved an apparently novel use (in the context of an operational ESL proficiency testing program) of the familiar regression model for the purpose of referencing scores on a norm-referenced ESL proficiency test to LPI performance, treated as a general, context-independent criterion measure.

The TOEIC Testing Context

The data employed in the study were not collected for ad hoc research purposes, but were generated in operational test-use settings in Japan (where the TOEIC was introduced in 1979), France, Mexico, and Saudi Arabia (countries in which the TOEIC has been introduced more recently). Detailed information regarding the TOEIC testing program and its operations is available elsewhere (e.g., ETS, 1985b, 1986a, 1986b, 1988: 8). For perspective, a general overview is provided below.

TOEIC Testing Programs

The TOEIC is used primarily by corporate clients outside the United States. The majority of TOEIC examinees are tested in places of work, or work-related ESL training, at the behest of employers (instructors), in group administrations of previously administered editions of the TOEIC, as part of the TOEIC Institutional Program (IP). In Japan and Korea, the TOEIC is also offered in three (3) national TOEIC Secure Program (SP) administrations annually, involving new forms of the TOEIC, for which individual preregistration is required.

In Japan, Korea, and several other countries, TOEIC-related assessment services are provided under the aegis of national TOEIC representative offices. In countries without national TOEIC representative offices, the TOEIC and TOEIC-related assessment services are obtained, by ad hoc arrangement with the TOEIC/ETS (Princeton) office, through the TOEIC International Corporate Program (ICP). ETS is responsible for test-development, scoring for SP administrations, and general oversight of TOEIC affairs worldwide. TOEIC programs are most highly developed in Japan where the TOEIC was introduced in 1979 at the request of the Japanese Ministry of International Trade and Industry (MITI). Currently, the majority of TOEIC corporate/institutional clients are located in Japan, as are about 80 percent of all TOEIC examinees--hundreds of thousands of examinees have been tested in Institutional Program (IP) and Secure Program (SP) test administrations under auspices of the TOEIC Steering Committee

in Japan. The TOEIC and TOEIC-related assessment services are also being used regularly, though to a lesser extent, in a number of other countries.

Characteristics of the Examinee Population

TOEIC examinees, worldwide, are likely to be

1. adult ESL users/learners who are relatively highly educated, typically at or beyond secondary-educational levels in national educational systems that provide formal instruction in English as a foreign language (EFL), and whose language-learning background is characterized by a core of academic EFL instruction, often with additional intensive ESL instruction;

2. employed or preparing for employment in ESL-essential positions, at home or abroad, with a business engaged in international business, commerce, or industry; hence directly or indirectly screened on pertinent employment-related cognitive, educational, personal, or other criteria, including English proficiency; and

3. tested in their places of work or work-related intensive ESL training, in administrations under the supervision of local TOEIC representatives or company-designated personnel.

TOEIC examinees in Japan, for example, are largely university-educated. They share a basic core of exposure to curriculum-embedded English-language instruction.[16]

Focus of the Present Study

The present study focuses primarily on data pertaining to the TOEIC testing context in Japan, where TOEIC/LPI data-sets have been generated for several years. All data-sets employed were derived from comprehensive operational ESL-assessments involving the concurrent use of the TOEIC and the LPI procedure, conducted in representative TOEIC-use settings by resident ESL professionals trained (and periodically "recalibrated") as LPI interviewers/raters in workshops conducted by TOEIC-ETS staff. Individual TOEIC-score data were available for general samples of Japanese TOEIC-SP examinees. Similar, but less extensive, TOEIC/LPI data-sets were also available for samples of examinees from representative TOEIC-use settings in three other countries: France (F), Mexico (M), and Saudi Arabia (S). These data-sets were generated in comprehensive ad hoc ESL assessments, also involving the joint use of the TOEIC and the LPI procedure, conducted by TOEIC/ETS staff members for corporate clients in those countries in 1987 and 1988. Due to the incipient nature of the testing programs in the countries involved, individual TOEIC-score data for general samples of French, Mexican, and Saudi examinees were not available for analysis.

Analytical Approach and Study Procedures

The TOEIC/LPI data for the samples described generally above were analyzed, using LPI performance as the criterion variable in the familiar regression model. On the basis of evidence and lines of reasoning developed in detail in the preceding section, it was assumed from the outset that the regression results would provide evidence needed to permit TOEIC users to make

statistically delimited interpretive inferences from TOEIC scores, about probable level of oral English proficiency in samples of ESL users/learners from the TOEIC testing contexts such as those represented in the study.

It was expected that this regression-based approach would generate useful interpretive guidelines for the TOEIC because ratings (scores) on the LPI criterion have direct "representational value," and because regression results, by definition, can be expected to indicate the extent to which the TOEIC scores share that "representational value." In other words, the regression results would contribute to the development of a defined "expectancy-set" about examinees' functional ability to use English based on their test scores.

An assessment was made of the level and pattern of concurrent correlation between the LPI criterion and TOEIC scores (LC, R, and Total [LC+R]) in the comparatively large Japanese TOEIC/LPI calibration sample, in the several non-Japanese samples, individually, in the total non-Japanese sample, and in the combined sample of Japanese and non-Japanese examinees.

It was hypothesized that the level of developed oral English proficiency (defined operationally by LPI performance) would be linked more closely to the level of developed English-language listening comprehension (indexed by TOEIC-LC) than to the level of developed English-language reading ability (TOEIC-R) in samples of educated ESL users/learners in representative TOEIC-use settings.

On logical/theoretical grounds, the ability to comprehend spoken English may be expected to affect performance in an interview situation, in which listening comprehension is measured semi-directly. This is not true in the case of reading ability. TOEIC-R clearly is an indirect measure of oral English proficiency.

A question that is of both theoretical and pragmatic interest has to do with whether a composite of LC and R scores tends to be more valid for predicting criterion (LPI) performance, than either of the two scores alone.

In analyzing the data, particular attention was given to evaluating the relative usefulness of three different TOEIC/LPI linkage equations, namely, (a) one equation based solely on the Listening Comprehension score, (b) a second equation based on TOEIC-Total (the simple sum of LC and R scores, informally weighted according to their standard deviations), and (c) a third equation specifying a "best-weighted" composite of LC and R, based on regression results in a particular sample. Equations were generated using TOEIC/LPI data for (a) the Japanese sample, (b) the total non-Japanese sample (that is, the French, Mexican, and Saudi examinees), and (c) the combined Japanese and non-Japanese samples.

Attention was focused first on an assessment of TOEIC/LPI relationships in samples of Japanese examinees. LPI ratings were regressed on TOEIC scores to (a) evaluate the degree and nature of association between functionally scaled LPI ratings (the criterion variable) and TOEIC-LC, TOEIC-R, and TOEIC-Total, and (b) develop equations for estimating LPI ratings from TOEIC scores in the Japanese testing context. Equations developed in the calibration sample

were used to estimate the distribution of criterion performance in general samples of Japanese TOEIC examinees.

An analysis was then made of TOEIC/LPI relationships in the French, Mexican, and Saudi samples, in the total FMS (non-Japanese) sample, and in the combined FMS and Japanese samples. This analysis was designed to assess the consistency of TOEIC/LPI linkage across samples from representative TOEIC-use settings in several linguistically diverse national TOEIC subpopulations. Data generated in the ad hoc assessments in Mexico and France, respectively, were used to conduct sub studies concerned with two areas not directly at issue in the study, namely, (a) the reliability of the LPI ratings employed and (b) the potential usefulness of self-assessments of oral English proficiency for research purposes. The latter substudy was suggested by the findings of Hilton et al. (1985).

Before focusing directly on the details pertaining to the foregoing lines of inquiry, it is important to provide a brief description of the TOEIC and its psychometric properties; also to elaborate briefly on the nature of the testing program in Japan, and to describe the role of direct proficiency assessment in the TOEIC testing program in Japan--that is, to describe factors associated with the development and maintenance of a cadre of ESL professionals trained in the LPI procedure, in representative TOEIC-use settings in Japan.

Characteristics of the TOEIC

The TOEIC is a multiple-choice, norm-referenced ESL proficiency test that provides measures of English language listening comprehension and reading abilities, respectively. (For an independent review and evaluation of the TOEIC, see Perkins [1987; 81-82]; see Woodford [1982] for developmental detail).

According to the Guide for TOEIC Users (ETS, 1986a: 1), reading items reflect the types of skills involved in comprehending types of ". . . materials that people in the business world use, including manuals, reports, forms, notices, advertisements, periodicals, and memoranda." The listening items are designed to measure understanding of spoken English in real-life situations.

Number-right raw-scores on the respective sections (listening and reading), each made up of 100 items, are translated into an arbitrarily defined standard-score scale with scores ranging from 5 to 495; a total score is derived simply by adding the two scaled section-scores. About two hours of actual testing time are involved.

The original form of the TOEIC was developed (in 1979) using items of appropriate difficulty for samples composed predominantly of university-educated adult Japanese nationals in or preparing for positions requiring the use of English as a second language (Woodford, 1982). ETS develops three different forms of the test each year. These forms are equated, through statistical linkage formulas, to assure comparability of scores across successive forms. Equating computations are carried out using data for samples of Japanese examinees who participate in regularly scheduled Secure Program (SP) test administrations (e.g., Angell, Gallagher, and Schneider, 1988). Computerized data files for SP administrations (offered only in

Japan and Korea, as of 1989) are maintained by ETS (Princeton) for purposes of test development and analysis.

Reliability coefficients for the two section scores in these equating samples tend to be in the mid-.90's; total score reliability typically is slightly higher than that for either section. Thus, the TOEIC provides a highly reliable basis for assessing individual and group differences in acquired English language listening comprehension and reading skills.

Evidence of Concurrent Validity

The test has face validity as a measure of reading and listening comprehension in English. Available empirical evidence (e.g., Woodford, 1982) suggests that the TOEIC-LC and TOEIC-R scores are correlated relatively strongly with corresponding LC and Reading Comprehension & Vocabulary scores on the Test of English as a Foreign Language (TOEFL), a test that is widely used to assess the English language skills of foreign-ESL students applying for admission to U.S. and Canadian colleges and universities (see, for example, ETS, 1985a). However, each of the two tests contains item types not found in the other.

Table 1 provides information regarding concurrent relationships between TOEIC and TOEFL scores in a sample of Japanese test takers. Test means were not reported for the TOEIC/TOEFL sample for which intercorrelations are shown in the table. Typical reliability coefficients for the two tests are also shown.

In a TOEIC-validation study (Woodford, 1982) involving data for a sample (N = 99) of Japanese examinees from the introductory (1979) test administration in Japan, TOEIC-LC scores were relatively highly correlated (about .80) with concurrent LPI ratings (based on interviews conducted by native-English speaking ESL professionals in Japan, trained especially for the study). Correlations at about this level were also reported for TOEIC-LC and/or TOEIC-R with concurrent direct measures of listening, reading, and writing that were developed ad hoc--direct listening and reading measures involved, for example, taped and written English stimuli, respectively, with questions and answers in Japanese.

The results of Woodford's (1982) study indicated a relatively high level of concurrent correlation between TOEIC scores and all the direct measures of proficiency, including the Language Proficiency Interview.

Introduction of the LPI Procedure in TOEIC-Use Settings in Japan

The interviewer/raters who generated the LPI ratings used by Woodford (1982) were recruited and trained especially for an ad hoc validity study. To assure the continued availability of a cadre of trained LPI interviewers/raters in the Japanese TOEIC-testing context, a TOEIC-ETS staff member conducted in Japan (in 1982) the first of a continuing series of workshops designed to provide training for conducting and rating interviews.

The participants in these workshops are native-English-speaking ESL professionals resident in Japan. They typically are responsible for conducting continuing, on-site intensive

Table 1
Concurrent Correlation Between TOEIC and TOEFL Score in a Sample of Japanese TOEIC Examinees

Variable	TOEIC			TOEFL			
	LC r	R r	Total r	LC r	S&WE r	RC&V r	Total r
TOEIC							
L	(.92)	.77	[.94]	.87	.74	.80	NA
R		(.93)	[.94]	.78	.85	.87	NA
Total			(.96)	NA	NA	NA	NA
TOEFL*							
LC				(.89)	.67	.68	[.86]
S&WE					(.86)	.78	[.92]
RC&V						(.90)	[.92]
Total							(.95)

Note: Entries in parentheses are estimated reliability coefficients in general examinee samples; entries in [brackets] are part-whole coeffients (TOEIC [ETS, 1980], TOEFL [ETS, 1985a]). TOEIC/TOEFL correlations were obtained in a sample of TOEIC examinees, who were asked to take a special administration of the TOEFL in Japan in 1979; means were not reported (ETS, 1982a; Woodford, 1982).

*The TOEFL sections are: Listening Comprehension, Structure and Written Expression, Reading Comprehension and Vocabulary.

programs of ESL training sponsored by corporations, or for conducting such programs in educational institutions. In addition to using the LPI procedure in their respective employment contexts, from time to time some of these specialists, by arrangement with the TOEIC Steering Committee in Japan, provide interview-assessment services under TOEIC auspices for individuals or groups of individuals.[17]

Members of the cadre of Japan-based ESL professionals involved in these TOEIC-related LPI workshops generated the TOEIC/LPI data-sets for the samples of Japanese examinees involved in the present study.

Source of TOEIC/LPI Data for Japanese Examinees

One data-set consisting of TOEIC scores and LPI ratings for 122 individuals was collected at the initiative of the TOEIC Steering Committee in 1985. Three additional data-sets (for a total of 163 individuals) were collected during the course of periodic, comprehensive ESL proficiency assessments involving the joint use of TOEIC and interviews that were conducted (in 1984, 1986, and 1987) by the English Department of the Institute for International Studies and Training (IIST), a graduate-level business school (and a regular institutional TOEIC subscriber).[18] The TOEIC/LPI-calibration sample thus consisted of what may be referred to as "TOEIC" and "IIST" subsamples.

The TOEIC subsample. The TOEIC subsample was selected and tested for the explicit purpose of evaluating concurrent TOEIC/LPI relationships in samples from a representative array of TOEIC-use contexts identified as corporations or other organizations in the Tokyo area that regularly use TOEIC services. On-site interviews (taped, and rated by at least two individuals) were conducted for previously tested TOEIC examinees in a number of corporate or ESL training sites; a few individuals were interviewed in the TOEIC office.

The IIST subsamples. The IIST, among other programs, offers a nine-month business-oriented training program conducted in Japanese, supplemented by intensive ESL instruction. The program consists of an eight-week intensive English course, a 14-week course in Area Studies and Basic Economics and another 14-week course in International Management and Economics. Trainees in the program typically are selected by a sponsoring company or governmental agency, not by the IIST. The trainees are predominantly male. An estimated 95 percent are university graduates; as such they typically have had a core of academic exposure to the study of English as a foreign language (see earlier note, p. 30).

The English-language needs of trainees vary: some are scheduled for overseas assignments or for ESL-essential jobs in Japan after program completion; ESL needs are less immediate for other trainees. The TOEIC and the LPI are used jointly so that sponsoring organizations will have "a recognized norm by which they can measure the English-language skills of their trainees" (Reilley, March 3, 1988, facsimile communication). TOEIC/LPI data-sets for three groups of IIST trainees evaluated toward the end of the ESL segment of the program in 1984, 1986, and 1987, respectively, were made available for use in the present study.

Analysis of TOEIC/LPI Relationships in Samples of Japanese Examinees

Means and standard deviations of the study variables are shown for the four Japanese samples in Table 2. They are labeled according to origin as TOEIC-85, IIST-84, IIST-86, and IIST-87. For perspective, comparable statistics are provided for a one-third sample of Japanese Secure Program examinees from the September 1987 administration (data from TOEIC-ETS files). The average TOEIC scores in the TOEIC/LPI samples were somewhat higher than those for SP examinees generally--Total-score means were 618 and 548, respectively.

Table 2

Summary Statistics for the Japanese Calibration Sample(s)

Sample	N	TOEIC scores LC M	LC SD	R M	R SD	Total M	Total SD	LPI* rating M	LPI* rating SD
JAPAN	285	316	83	302	77	618	151	1.86	0.67
TOEIC-85	122	311	92	300	78	611	160	1.89	0.72
IIST-84	66	313	77	295	76	608	145	1.78	0.66
IIST-86	55	329	73	310	72	640	137	1.87	0.59
IIST-87	42	315	76	309	83	624	151	1.93	0.61
SP-87	3,558**	288	91	260	91	548	172	N.A.***	

*LPI level (See Appendix A for detailed descriptions of levels.)

5 Function is equivalent to that of an educated native speaker.
4 Able to tailor language to fit audience, counsel, persuade, negotiate, represent a point of view, and so on.
3 Can converse in formal and informal situations, describe in detail, offer supported opinions, and so on.
2 Able to fully participate in casual conversations, can speak in extended discourse, express facts, give instructions, describe, report, and provide narration about current past and future activities.
1 Can create with the language, ask and answer questions, participate in short conversations.
0 No functional ability.

** Randomly selected examinees from the September, 1987, SP administration (data in TOEIC-ETS files).

*** To be estimated.

It is apparent that these numbers do not convey any information about how well Japanese examinees are able to function in English. By inference, examinees in the TOEIC/LPI sample are likely to have more functional ability to use English than those in the general SP sample. On the other hand, the fact that the mean LPI rating for the calibration sample was approximately at LPI Level 2 conveys some directly interpretable information.

For example:

o Interviewees at LPI Level 2 able to participate fully in casual conversations, can speak in extended discourse and express facts, give instructions, describe, and provide narration about current past and future activities (see the detailed description of this level in Appendix A).

The four subsamples appear to be similar with respect to overall patterns of performance on the study variables. This general impression is confirmed by the results of a multiple discriminant analysis (MDA), not reported in detail, indicating that the joint distributions of TOEIC scores and LPI ratings for the groups were not significantly different.[19]

Regression Results

Table 3 shows intercorrelations of the study variables in the combined (Japan) sample and in the four subsamples. Selected results of multiple regression analyses in the respective samples are shown in the upper right portion of the table--that is, standard partial regression coefficients (beta weights) and multiple correlation coefficients that were obtained when LPI ratings were regressed on TOEIC LC and R scores in the respective samples. Several trends are noteworthy.

o TOEIC-LC was more closely associated with the LPI criterion than was TOEIC-R, as hypothesized, except in the IIST-1984 data-set.

o The coefficient for TOEIC-LC was comparable to that for TOEICTotal, and only slightly lower than the multiple correlation coefficient reflecting the best-weighted combination of the LC and R.[20]

o The Total score (LC plus R), was about as closely related to the LPI criterion as was the best weighted combination of LC and R (com-pare multiple correlations, shown in the rightmost column of Table 3, with simple correlations for TOEIC-Total.

Consistency of LPI-estimation from TOEIC scores.

To assess stability of fit between observed and estimated LPI ratings across the four samples, a residual analysis was performed. Using data for the combined Japan sample, LPI ratings were regressed on TOEIC-LC, TOEIC-LC and TOEIC-R, and TOEIC-Total. The resulting regression equations were used to compute three different criterion estimates and the corresponding residual values (that is, observed minus estimated LPI ratings) for each individual. Means and standard deviations of the residual values were then computed for each of the subsamples. Results of a one-way analysis of variance (ANOVA) of the residuals, shown in Table 4, indicate a very close fit between the observed and estimated values across the four samples, regardless of the equation employed. In all instances, the mean difference between estimated and observed LPI ratings was less than 0.1 on the 11-point LPI scale. The residual standard deviations were comparable to the standard error of estimate

Inferring LPI Performance from TOEIC Scores

On the basis of the foregoing findings, data for the four subsamples were combined for analyses designed to highlight the degree of fit between estimates of LPI rating based on each of the two TOEIC-score equations, and actual LPI rating throughout the range of TOEIC scores represented in the study sample.

Table 3

Intercorrelations of Variables in the Calibration Sample(s), and Results of Multiple Regression Analysis (Beta Weights and Multiple Correlation Coefficients)

Variable/Sample		Simple Correlations				Regression results* Beta weights		(R)
	N	LC	R	TOTAL	LPI	LC	R	
LC -JAPAN	285	--	.79	(.95)	.75	.55	.26	.77
TOEIC-85	122	--	.78	(.95)	.79	.58	.26	.80
IIST-84	66	--	.80	(.95)	.67	.33	.42	.71
IIST-86	55	--	.81	(.95)	.80	.82	-.01	.80
IIST-87	42	--	.80	(.94)	.73	.49	.30	.76
R -JAPAN	285	--	--	(.94)	.69			
TOEIC-85	122	--	--	(.94)	.72			
IIST-84	66	--	--	(.95)	.68			
IIST-86	55	--	--	(.95)	.65			
IIST-87	42	--	--	(.95)	.70			
TOTAL-JAPAN	285	--	--	--	.76			
TOEIC-85	122	--	--	--	.80			
IIST-84	66	--	--	--	.71			
IIST-86	55	--	--	--	.76			
IIST-87	42	--	--	--	.75			

Note. Coefficients in parentheses reflect part-whole correlation.

* The data shown in the upper right portion of the table are standard partial regression (beta) weights for TOEIC-LC and TOEIC-R, and the multiple correlation coefficient (R), obtained for analyses in the total (Japan) sample and in the respective subsamples.

Table 4

Results of Residual Analysis for the Japanese Calibration Subsamples Using Several Linkage Equations

		LPI criterion estimated by					
		LC*		LC & R**		TOTAL***	
Sample	(N)	Mean resid.	SD resid.	Mean resid.	SD resid	Mean resid.	SD resid.
TOTAL	285	.00	.44	.00	.43	.00	.43
TOEIC-85	122	.05	.44	.04	.40	.04	.43
IIST-84	66	-.07	.50	-.06	.47	-.05	.47
IIST-86	55	-.07	.40	-.07	.36	-.07	.38
IIST-87	42	.04	.42	.05	.40	.04	.41
F-Ratio		1.80		1.52		1.29	
Probability		0.14		0.21		0.27	

*LPI = (.006067*LC) + .049348 [R = .75]
**LPI = (.004401*LC) + (.002266*R) .208272 [R = .77]
***LPI = (.003376*Total)+ .220179 [R = .76]

Table 5 shows, for designed TOEIC-Total intervals (upper section) and TOEIC-LC intervals (lower section), (a) the number of examinees in the calibration sample, (b) the LPI level expected for individuals at the midpoint of each interval, based on the regression equation, and (c) the mean and standard deviation of the observed LPI ratings for examinees in each interval. The standard error of estimate in each case was approximately .50 (.45) on the LPI scale (see the "Actual S.D." values in the last column of Table 5).

Fit between actual and estimated LPI means at various TOEIC score levels is shown in Figure 4a (for TOEIC Total) and Figure 4b (for TOEIC-LC). The points plotted in the two figures correspond to the interval-mean LPI ratings of individuals in the calibration sample.[21] In each figure, the points conform closely to the line specified by the regression equation, throughout the range of TOEIC scores represented in the calibration sample--100 or higher for LC, 200 or higher for Total. The horizontal lines are spaced at .5 intervals that correspond, approximately, to the respective standard errors of estimate.

Table 5

Estimated and Observed LPI Levels Associated with Designated Levels of Performance on TOEIC Total and TOEIC Listening Comprehension, Respectively

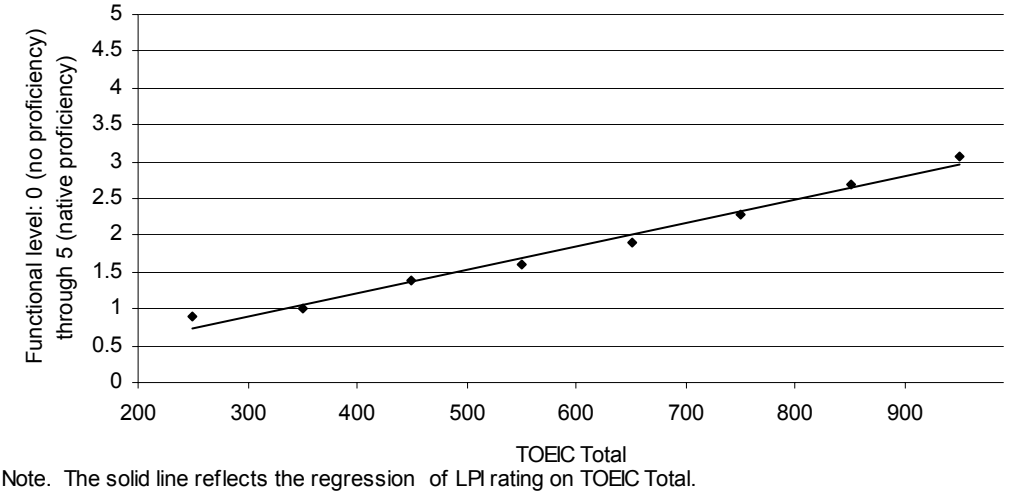
TOEIC interval	Midpoint	N	Mean LPI rating		
			Estimated	Actual	Actual S.D.
TOEIC-Total interval					
200-299	250	(5)	.62	.90	.42
300-399	350	(16)	.96	1.00	.37
400-499	450	(41)	1.30	1.38	.42
500-599	550	(68)	1.64	1.59	.38
600-699	650	(67)	1.97	1.89	.44
700-799	750	(48)	2.31	2.29	.49
800-899	850	(31)	2.65	2.68	.47
900+	950	(9)	2.99	3.06	.53
TOEIC-LC interval					
100-149	125	(4)	.71	.63	.25
150-199	175	(18)	1.01	1.19	.39
200-249	225	(37)	1.32	1.34	.46
250-299	275	(62)	1.62	1.58	.38
300-349	325	(72)	1.92	1.85	.45
350-399	375	(38)	2.23	2.11	.47
400-449	425	(37)	2.53	2.65	.53
450+	475	(17)	2.83	2.85	.49
Total sample		(285)	1.86	1.86	.67

Note. Estimated LPI values (LPIest) are based on the following equations:

$LPIest_{(Total)} = (.003376*Total) - .220179$

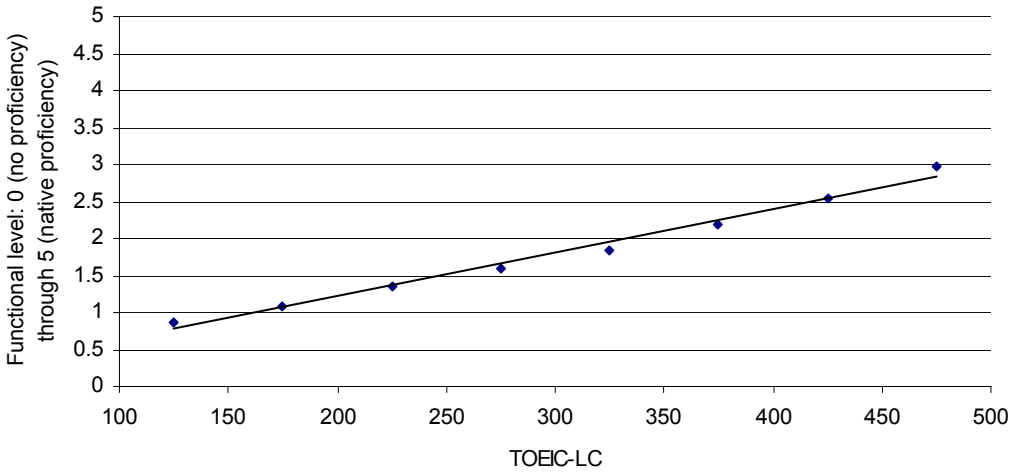
$LPIest_{(LC)} = (.006067*LC) - .049348$

Figure 4a. Fit between actual LPI means and means estimated from TOEIC Total, by Total-score intervals: Data for the calibration sample (N=285)



Note. The solid line reflects the regression of LPI rating on TOEIC Total.

Figure 4b. Fit between actual LPI means and means estimated from TOEIC-LC, by LC-score interval: Data for the calibration sample (N=285)



Note. The solid line reflects the regression of LPI rating on TOEIC-LC.

Average LPI expectancy increases directly with TOEIC performance. For example, an average LPI rating at Level 1 is expected for examinees with Total scores of about 350, an average average rating of Level 2 for those with Total scores at the 650 level, and an average rating of Level 3 for those scoring 950. The amount of variability expected in the LPI ratings at each score level is defined, statistically, by the standard error of estimate.

Tables 6.1 and 6.2 provide more comprehensive information. These are expectancy tables that show the actual distributions of LPI ratings (in percent) by TOEIC-score intervals, for TOEIC-Total and TOEIC-LC, respectively. The distribution of LPI ratings for the total calibration sample is also shown in the tables. From the tables it may be inferred, for example, that 90 percent or more of examinees with Total scores of 900 or higher or LC scores of 450 or higher earned LPI ratings of at least Level 2+; that the modal LPI rating for examinees in the 600-695 range (average of about 650) was Level 2; and so on.

Judging from the expectancy tables and previously considered findings, it appears that inferences about examinees' LPI performance (level of oral English proficiency) are likely to be equally valid, whether based on TOEIC-LC or on TOEIC-Total, and that inferences about criterion performance from TOEIC-Total, in turn, are comparable statistically to those based on complex, regression-weighted composites of LC and R.

This outcome is understandable statistically because TOEIC-LC is very highly correlated with TOEIC-Total (r's of about .95, artifactually inflated due to part-whole [self] correlation). In addition, the LC and R scores themselves are closely related (r's of about .80). The outcome is theoretically consistent because the ability measured by TOEIC-LC (to comprehend spoken English) is an integral aspect of the functional ability (to comprehend and produce utterances in English) that is assessed in the Language Proficiency Interview; reading skills, on the other hand, are not assessed directly or semi-directly in the interview situation.

At the same time, it should not be overlooked that TOEIC-R scores were relatively strongly correlated with the LPI criterion in the calibration sample (r =.69). This means that in the hypothetical absence of LC scores, very useful inferences about LPI performance could be drawn from examinees' scores on the TOEIC Reading section only--an indirect measure of oral English proficiency. To emphasize this point, trends in TOEIC-R/LPI relationships are shown in Table 6.3 and Figure 4c. These trends strongly parallel those for LC/LPI relationships shown above (in Table 6.2 and Figure 4b).[22]

Estimating the Distribution of Criterion Behavior
In General Samples of Japanese Examinees

It may be recalled (from Table 2) that the TOEIC-Total mean for the calibration sample was 618 (SD = 151), and the TOEIC-LC mean was 316 (SD = 83, as compared to means of 548 (Total)

Table 6.1

Relationship between TOEIC Total Score and LPI Rating in English: Japanese Sample

	Percent with LPI rating							
TOEIC	0+	1	1+	2	2+	3	>3	Total
Total								
900+					33	33	33	(9)
800-895				16	45	29	10	(31)
700-795			12	38	31	16	2	(48)
600-695		9	22	57	8	4		(67)
500-595		19	46	34	2			(68)
400-495	7	24	56	10	2			(41)
300-395	25	50	25					(16)
200-295	40	40	20					(5)
Total	3.2	13.7	28.1	30.9	13.7	8.1	2.5	
(N)	9	39	80	88	39	23	7	(285)

Note. This table is based on data for 285 TOEIC examinees tested in Japan.

Table 6.2
Relationship between TOEIC Listening Comprehension Score and LPI Rating in English: Japanese Calibration Sample

	Percent with LPI rating							
TOEIC	0+	1	1+	2	2+	3	> 3	Total
LC								
450+				6	41	35	18	(17)
400-445			3	19	38	30	11	(37)
350-395		3	18	42	29	8		(38)
300-345		11	21	58	6	4		(72)
250-295	2	13	56	26	3			(62)
200-245	8	35	41	14	3			(37)
150-195	11	44	39	6				(18)
< 150	75	25						(4)
Total	3.2	13.7	28.1	30.9	13.7	8.1	2.5	
(N)	(9)	(39)	(80)	(88)	(39)	(23)	(7)	(285)
LPI	0+	1	1+	2	2+	3	> 3	

Note. Data for 285 TOEIC examinees tested in Japan.

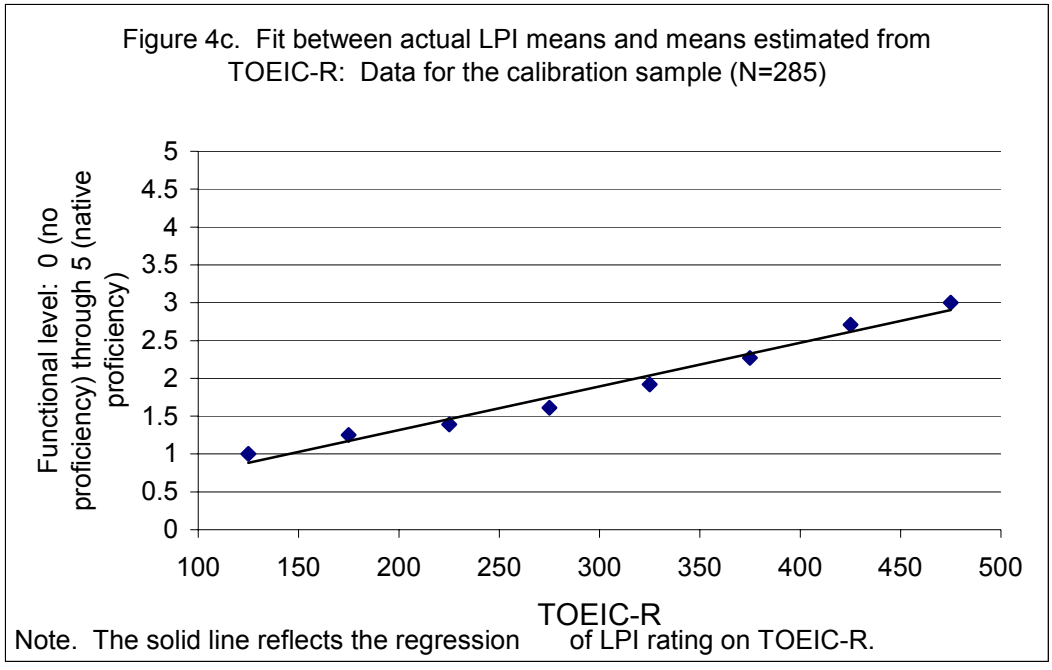
Table 6.3

Relationship between TOEIC Reading Score and LPI Rating in English: Japanese Calibration Sample

	Percent with LPI rating							
TOEIC Reading	0+	1	1+	2	2+	3	>3	Total (N)
450+						100#		(1)
400-445			3	19	29	32	16	(31)
350-395		7	7	33	36	16	2	(58)
300-345		8	27	47	12	5	2	(58)
250-295		15	48	37				(62)
200-245	7	27	48	18				(44)
150-195	19	38	25	12	6			(16)
< 150	33	33	33					(9)
Total (N)	3.8 (15)	11.5 (45)	26.0 (102)	31.8 (125)	14.2 (56)	10.2 (40)	2.6 (10)	(393)

* These are joint TOEIC/LPI data for 285 Japanese TOEIC examinees.

Based on a single case.



and 288 (LC) for a general sample of Secure Program examinees tested in September, 1987. Secure Program (SP) examinees are more highly selected than examinees tested in the TOEIC Institutional Program (IP).

According to information provided by the TOEIC Steering Committee in Japan, means for IP examinees generally are approximately 200, 200, and 400 for LC, Reading, and Total, respectively. As indicated earlier, individual score data were not available for IP examinees.

To provide perspective on the probable distribution of functional ability to use English in face-to-face conversation in the general Japanese-examinee population, LPI ratings were estimated from TOEIC Total for a sample of examinees from the TOEIC SP administration conducted in Japan in September 1987; estimates were also made of the mean LPI rating for the IP population.[23]

The distribution of estimated LPI levels for the SP sample is shown in Figure 5. The LPI mean was 1.6 (Level 1+) and the standard deviation was .5, corresponding to a TOEIC-Total mean of 548 and standard deviation of 172.

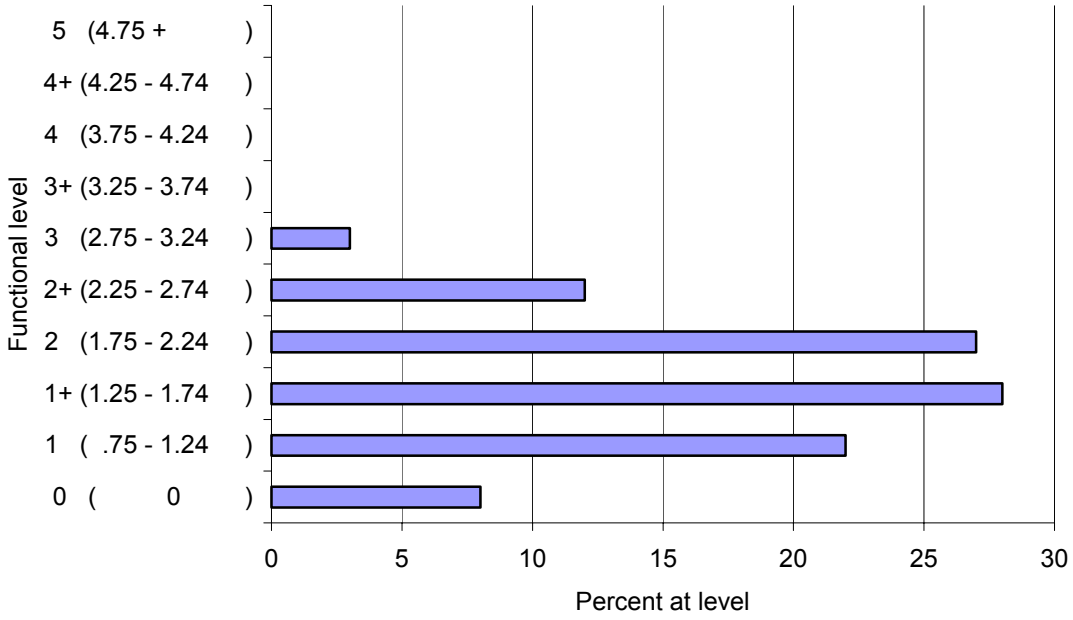
Based on the mean Total score of 400 reported for IP examinees, the mean estimated LPI performance for this sample is approximately at Level 1 (estimated mean = 1.13). Given a standard error of estimate of approximately .5, the majority of Japanese IP examinees probably are functioning conversationally at or below Level 1+ (1.5).8

Taking at face value the findings reflected in Figure 5, and considering that these findings do not reflect data for the lower-scoring IP examinees, certain general inferences as to the distribution of criterion behavior (LPI performance) in the TOEIC examinee population in Japan are warranted. For example:

1. Relatively few Japanese TOEIC examinees (SP and IP combined) are likely to be functioning conversationally at or higher than LPI Level 3.

2. The typical level of developed ESL conversational ability in the general TOEIC examinee population in Japan (combined IP and SP) is about as described for Level 1; the majority probably are functioning approximately between Level 0+ (.5) and Level 2+ (2.5).[24]

Figure 5. Distribution of estimated levels of oral ESL proficiency for a sample of Japanese TOEIC Secure Program examinees (in percent)



Note. See Appendix A for detailed descriptions of FSI/ILR speaking levels characterized briefly below:

FSI/ILR

Level 5 Functions equivalent to an educated native speaker.

Level 4 Able to tailor language to fit audience . . . on all topics pertinent to professional needs.

Level 3 Can converse in formal and informal situations . . .deal with unfamiliar topics . . . offer supported options.

Level 2 Able to participate fully in casual conversations, speak in extended discourse, express facts, give instructions, report, add provide narration about current, past and future activities.

Level 1 Can create with the language, ask and answer questions, participate in short conversations.

Level 0 No functional ability.

Section IV. STABILITY OF TOEIC/LPI RELATIONSHIPS IN SAMPLES FROM DIVERSE TOEIC TESTING CONTEXTS

The findings reviewed above provide evidence regarding the pattern of concurrent correlations between TOEIC scores and LPI ratings in several samples from the majority (Japanese) examinee subpopulation, the average level and range of behaviorally defined LPI performance that may be expected of Japanese examinees who present particular TOEIC scores, and the probable distribution of criterion performance in the TOEIC testing context in Japan.

Questions naturally arise as to whether TOEIC/LPI relationships are consistent for ESL users/learners likely to be tested with the TOEIC in other countries. It was possible to conduct analyses of the consistency of TOEIC/LPI relationships across diverse samples, using data available for examinees from TOEIC-use settings in France, Mexico, and Saudi Arabia--that is, TOEIC/LPI data-sets for samples of employees in ESL-essential jobs in the Paris office (N = 56) and the Mexico City office (N = 42) of an international accounting firm, and Saudi employees (N = 10) of an international petroleum corporation.[25]

Although small, these samples of educated, adult ESL/users learners were from representative TOEIC-use settings--places of work or work-related ESL training that are generally similar in nature from country to country. Data for general samples of TOEIC examinees from France, Mexico, and Saudi Arabia were not available for analysis.

TOEIC/LPI Correlations in Diverse Samples

Table 7 shows the observed pattern of concurrent TOEIC/LPI correlations for the French (F), Mexican (M), and Saudi (S) samples, individually, the total FMS sample (N = 108), the combined FMS and Japanese samples (N = 393), and the Japanese calibration sample (N = 285). Means and standard deviations are also shown.

From Table 7, it is apparent that the general pattern of concurrent TOEIC/LPI relationships was similar across all the samples. Coefficients were somewhat lower in the French than in the Mexican and Saudi samples, consistent with differences in TOEIC-score variability. Standard deviations were larger in the Saudi and Mexican samples than in either the French or the Japanese sample. The coefficient for TOEIC-LC typically was larger than that for TOEIC-R, and comparable to the coefficient for TOEIC-Total. In the FMS total sample, the coefficient for LC was slightly higher than that for Total (with the reading component).

In the total FMS sample, when LPI was regressed on LC and R treated as independent predictors (results not shown in Table 7), the resulting multiple correlation coefficient (.744) was essentially identical to the simple LC/LPI correlation (r = .7439). In a similar analysis for the combined FMS and Japanese samples (N = 393), the best-weighted LC+R composite yielded a multiple correlation coefficient (R = .753) that was only very slightly higher than the simple LC/LPI and Total/LPI correlations shown in Table 7 (r = .745 in both instances).

Table 7
Data on Stability of TOEIC/LPI-Criterion Relationships Across Samples from Different TOEIC-Use Contexts

Sample	N	Correlation with LPI			Means and standard deviations							
		List r	Read r	Total r	LPI M	LPI SD	List M	List SD	Read M	Read SD	Total M	Total SD
France-87(F)	56	.62	.58	.65	2.30	.64	428	74	389	48	817	113
Mexico-87(M)	42	.78	.70	.76	1.71	.62	262	106	237	104	499	204
Saudi-87 (S)	10	.85	.86	.87	1.95	.93	304	107	184	114	489	217
(FMS total)	108	.74	.67	.73	2.04	.71	352	120	311	115	663	229
(Japan total)	285	.75	.69	.76	1.86	.67	316	83	302	77	618	151
Combined	393	.74	.68	.74	1.91	.68	325	96	305	89	630	177

Consistency of LPI Estimation Across Diverse Samples: A Residual Analysis

These correlational findings indicate that within the several nationally defined samples, and in the two nationally and linguistically heterogeneous "general" samples--that is, the total FMS sample, and the combined study sample--LPI-criterion performance varied more closely with the TOEIC-LC than with TOEIC-R, and the coefficient for LC only was comparable to that for TOEIC-Total. However, evidence of consistency in patterns of concurrent TOEIC/LPI correlations, alone, does not shed direct light on a question that is of considerable theoretical as well as practical interest:

> Will ESL users/learners (of the type likely to be taking the TOEIC) who present particular TOEIC scores tend to exhibit about the same average level of LPI performance, regardless of national-linguistic origin?[26]

Evidence bearing on this question was obtained by analyzing differences across the four national samples in mean residuals associated with three sets of regression-equations for estimating LPI, namely, Set A (estimates from a weighted composite of LC & R), Set B (estimates from TOEIC-Total), and Set C (estimates from TOEIC-LC only), each set including equations reflecting data for three different TOEIC/LPI "calibration samples:" the FMS sample (N = 108), the combined FMS and Japanese samples (N = 393), and the Japanese sample (N = 285).

Nine residual values, one associated with each of the nine linkage equations, were computed for each individual in the study sample.[27] Mean residuals for the four application

samples (the French, Mexican, Saudi, and Japanese samples) and the three calibration samples (FMS total, Japan, Combined) are shown in Table 8. Results of one-way analysis of variance tests of differences in mean residuals associated with the respective linkage equations are also shown.

Figure 6 is a plot of the mean residuals shown in Table 8. By reference to the figure a general evaluation may be made of trends in the relative size (in absolute value) of the residuals--the smaller the mean residual, the better the fit between average level of criterion performance and average level estimated from a particular regression equation. Differences in mean residuals, though statistically significant in most instances, were comparatively small in absolute magnitude--that is, less than /.25/ on the LPI scale (0-5).

Mean residuals associated with TOEIC-LC (Set C) equations typically were smaller than those associated with equations involving composites of LC and R--that is, either equations involving TOEIC-Total (Set B) equations reflecting "best-weighted" composites of LC and R (Set A).

Except in the case of the Saudi sample, the mean residuals associated with composite-score equations (Set A or Set B) were generally similar to the mean residuals associated with LC equations (Set C). For example, in the Japanese, French, and Mexican application samples, mean LC-related residuals ranged, in absolute value, between /.00/ and /.25/ (the larger means were associated with Japanese-based calibration equations applied in the French and Mexican samples), indicating relatively close agreement.[28] However, in the Saudi sample, mean residuals for Set A and Set B equations, all influenced by TOEIC-R, were considerably larger in most instances (ranging up to /.52/).[29]

For the Saudi sample, the TOEIC-LC mean (304) was considerably greater than the TOEIC-R mean (184), whereas these two means were not so divergent in the general calibration samples (e.g., 325 versus 305 in the combined FMS and Japanese samples). In the same (Saudi) sample, the LPI mean (1.95) was consistent with the higher LC mean rather than with either the low R mean or the R-influenced Total score. Although the Saudi sample is small, based on evidence from the TOEFL testing context, t here is reason to believe that a pattern of higher average performance on measures of English-language listening comprehension than on measures of reading is characteristic of educated Saudi ESL users/learners.[30]

It is noteworthy that in a sample with quite divergent means on TOEIC-LC and -R, indicative of differential levels of development of the corresponding English-language skills, the average level of LPI-assessed oral English proficiency was indexed more accurately by average level of developed listening comprehension than by level of developed reading ability. This is especially interesting in view of the fact that the several measures were highly intercorrelated in the sample--TOEIC/LPI correlations were in the mid-80's, for example.

Figure 7a shows the regression of LPI rating on TOEIC-LC in the combined sample (N = 393); trends in LPI means by Total-LC interval are shown for Japanese examinees (broken line)

Table 8

Mean Residuals for Study Samples in Analyses Involving LPI as Estimated from (a) Best-Weighted Composites of LC and R, (b) TOEIC Total Score, and (c) TOEIC-LC only, Using Equations Developed in Different "Calibration Samples"

"Applica-tion" sample	N	Set A Wtd LC + R Calibration sample FMS	JAPAN	Comb	Set B Total score Calibration sample FMS	JAPAN	Comb	Set C LC only Calibration sample FMS	JAPAN	Comb
France	56	-.08	-.25	-.16	-.09	-.23	-.15	-.07	-.24	-.15
Saudi	10	.15	.40	.30	.30	.52	.44	.12	.15	.15
Mexico	42	.07	.23	.17	.05	.25	.18	.07	.18	.14
Japan	285	-.02	.00	-.00	.03	.00	-.01	-.02	.00	.00
Combined	393	-.02	-.00	.00	-.05	.01	.00	-.01	-.01	.00
FMS total	108	.00	-.01	.01	.00	.02	.03	.00	-.04	-.01
F-ratio		1.3	12.9	6.0	2.9	14.3	7.7	1.0	7.8	3.9
Prob		.27	.00	.00	.03	.00	.00	.38	.00	.01

Note. Underscoring indicates that the mean is expected to be zero because the estimation equation involved is based on data for the corresponding "calibration sample" (the sample in which the regression equation involved was developed).

and for FMS examinees (dotted line). Figure 7b shows comparable trends involving TOEIC Total. In both samples, LPI performance tended to conform more consistently to expectation (the combined-sample regression line) based on TOEIC-LC than to expectation based on TOEIC-Total. A summary of evidence regarding the relationship between TOEIC-LC and LPI ratings in the combined sample is provided in Figure 8 and Table 9. These displays are comparable to those shown earlier for the Japanese calibration sample (see, for example, Figure 4b and Table 6.2, above.

Figure 6. Plot of mean residuals by sample as a function of TOEIC score(s) used for estimation and the calibration sample involved (data from Table 9)

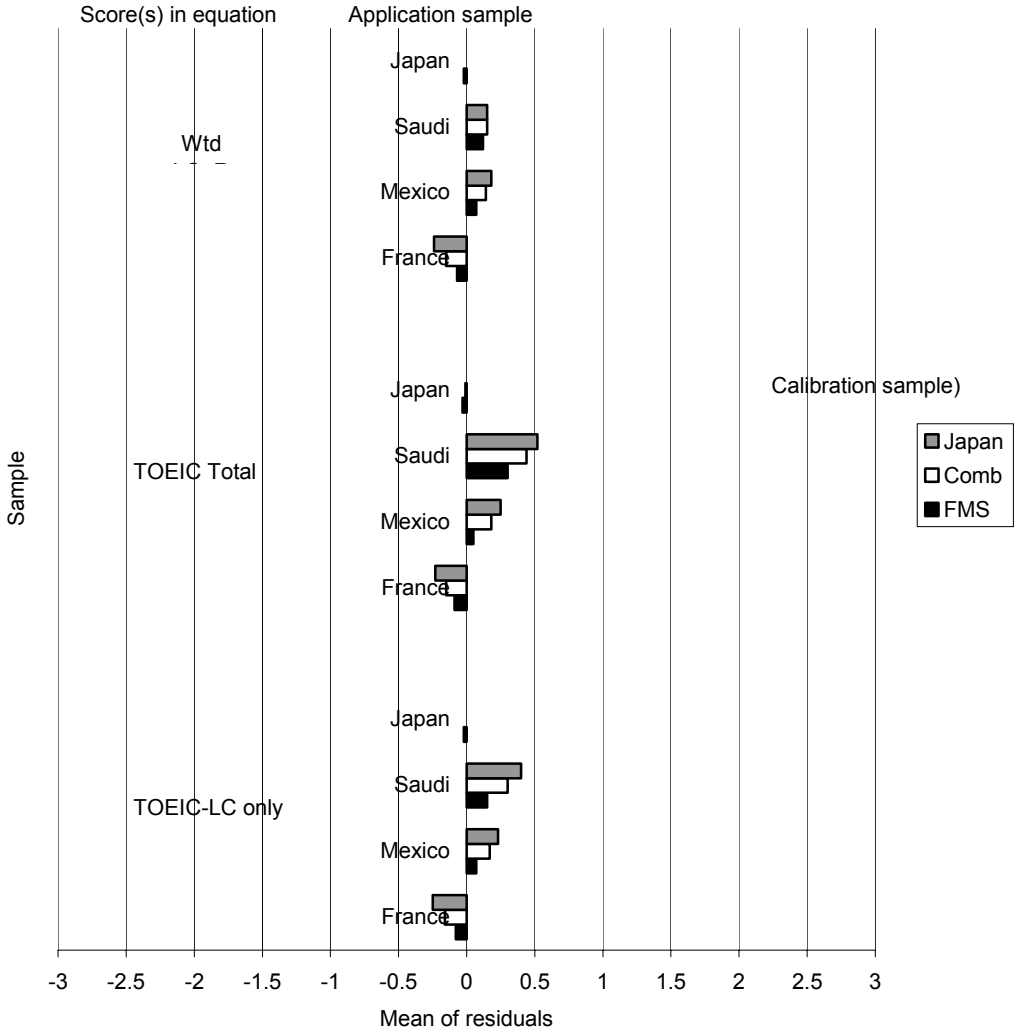


Figure 7a. Regression of LPI rating on TOEIC-LC in the combined sample (N=393), with plot of mean rating by LC-score interval for the TOEIC-Japan and TOEIC-FMS samples

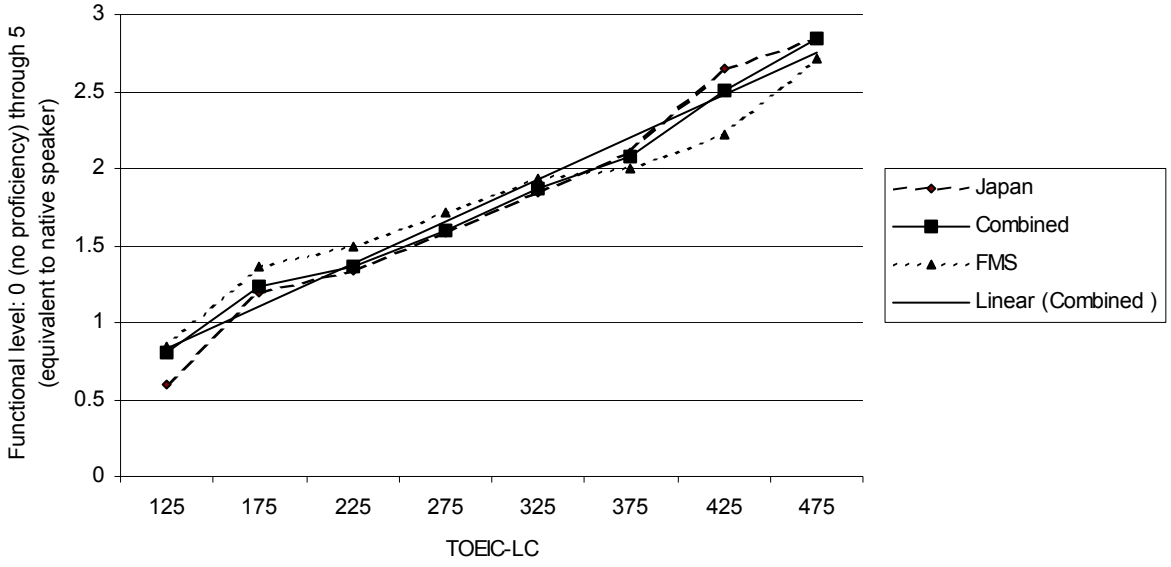


Figure 7b. Regression of LPI rating on TOEIC-Total in the combined sample (N=393), with plots of actual values for the TOEIC-Japan and TOEIC-FMS samples

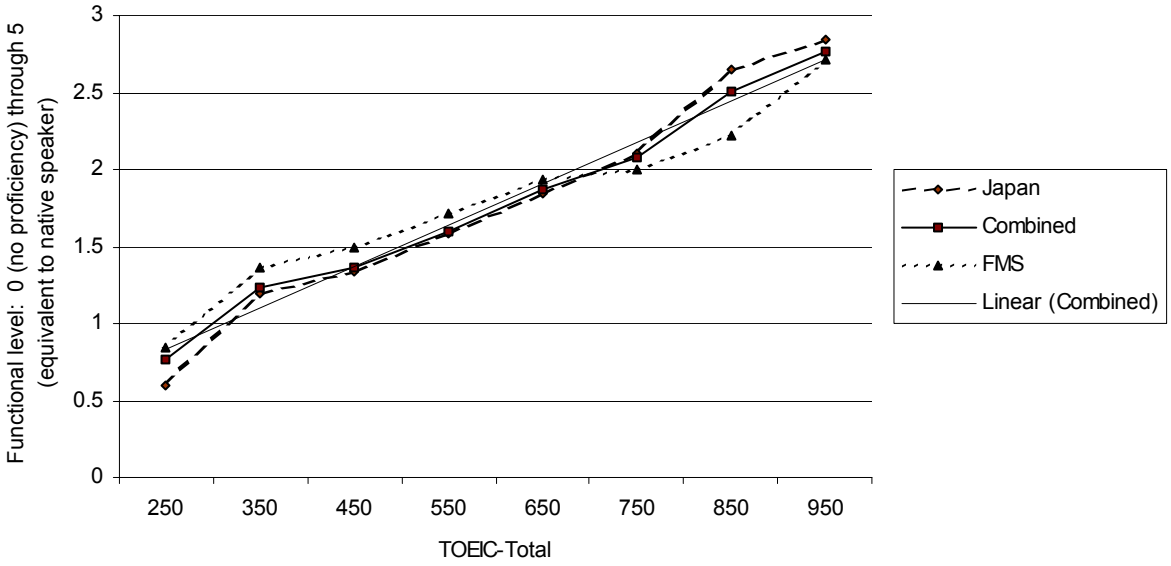
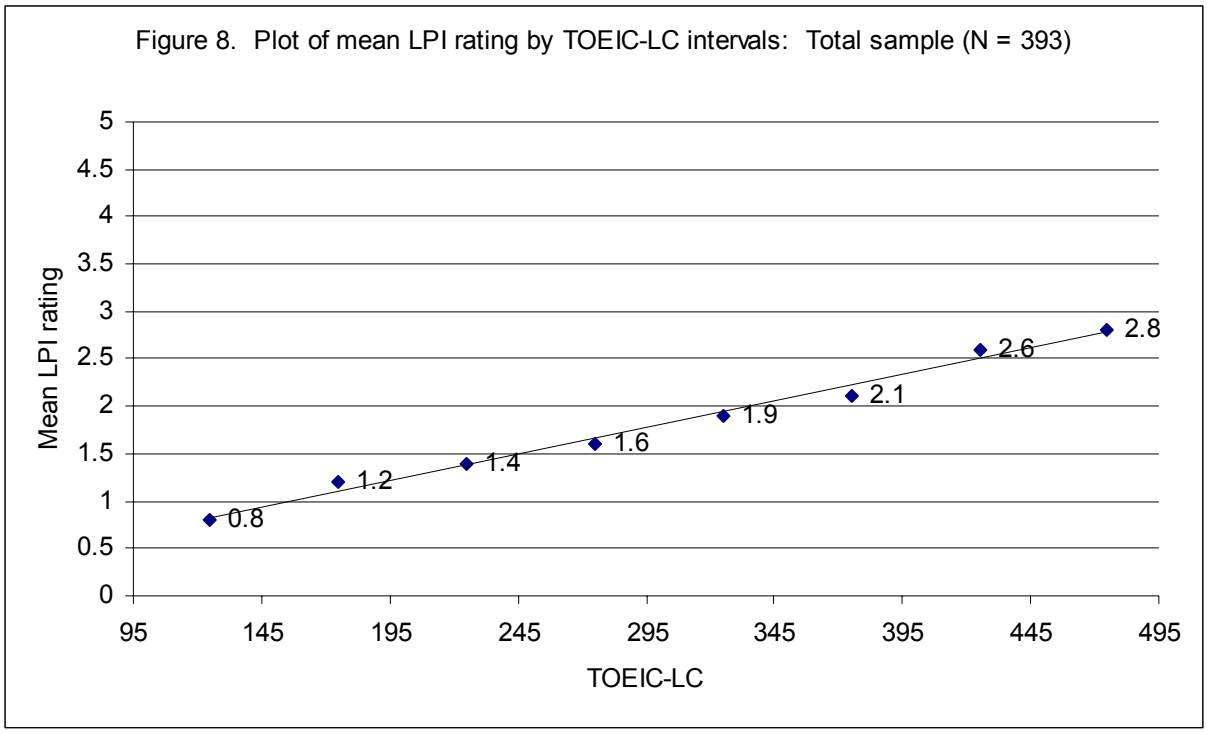


Table 9

Relationship between TOEIC Listening Comprehension Score and LPI Rating in English: Combined Sample*

Percent with LPI rating TOEIC-LC interval	0+	1	1+	2	2+	3	>3	Total (N)
450+				17	30	40	13	(47)
400-445			9	21	36	26	8	(53)
350-395		4	18	45	26	8		(51)
300-345		9	20	61	6	3		(88)
250-295	3	11	55	26	6			(73)
200-245	7	33	42	16	2			(45)
150-195	12	40	36	12				(25)
< 150	64	18	18					(11)
Total (N)	3.8 (15)	11.5 (45)	26.0 (102)	31.8 (125)	14.2 (56)	10.2 (40)	2.6 (10)	(393)

* These are joint TOEIC/LPI data for 393 TOEIC examinees tested in Japan (N = 285), France (N = 56), Mexico (N = 42), and Saudi Arabia (N = 10).



Section IV: TOEIC/LPI RELATIONSHIPS--FINDINGS, CONCLUSIONS, AND SUGGESTED DIRECTIONS FOR FURTHER INQUIRY

This study was undertaken to develop and evaluate regression-based guidelines for making inferences from (a) scores on the TOEIC, about (b) level of ability to use English in face-to-face conversation (indexed by performance in Language Proficiency Interviews), for (c) examinees in samples of ESL users/learners from the TOEIC testing context, using (d) data generated during the course of operational ESL assessments involving the joint use of TOEIC scores and the LPI procedure in diverse TOEIC-use settings. The findings reflect actual experience in representative test-use setting in Japan, and in France (F), Mexico (M), and Saudi Arabia (S)--countries in which the TOEIC has been introduced more recently.

Overview and Evaluation of Findings

Performance in the Language Proficiency Interview was strongly and consistently associated with TOEIC performance not only in the comparatively large TOEIC/LPI-calibration sample from the majority (Japanese) TOEIC test-taking subpopulation, but also in samples of examinees from three national subpopulations in the larger TOEIC testing context.[31]

Trends in TOEIC/LPI relationships observed in these samples reflect patterns of association that plausibly can be expected to hold in similar samples from the corresponding national TOEIC subpopulations and, by inference, in similar samples from the larger TOEIC testing context. Study findings are reviewed and evaluated in some detail, below, to highlight the evidentiary and theoretical foundation for this assertion, and for related conclusions and interpretive generalizations about the findings.

Consistent Pattern of Concurrent TOEIC/LPI Correlation

There was a consistent pattern of concurrent correlation between TOEIC scores (LC, R, and Total), and level of functional ability to use English in face-to-face conversation (LPI performance) in the study sample. The pattern was essentially as described below:

1. TOEIC-LC/LPI correlations (typically in the mid-70's) were somewhat higher than TOEIC-R/LPI correlations (typically about .70).

2. Simple correlations between the Total score (with the reading component) and the LPI criterion were about about the same as the LC/LPI coefficients--very slightly lower in some instances.

3. When LPI performance was regressed on LC and R (treated as a battery of predictors) in the Japanese, FMS (total non-Japanese) sample, and the combined FMS and Japanese samples, respectively, the resulting multiple correlations (uncorrected for shrinkage) were only very modestly larger than the simple correlation for TOEIC Total, or TOEIC-LC only.

Functional Linkage Suggested Between Listening Comprehension and Oral Language Proficiency

Viewed from a theoretical perspective, evidence of consistently higher criterion-related validity for TOEIC-LC than for TOEIC-R, and lack of improvement in prediction when the R score is added to the LC score, suggests a strong underlying functional linkage between the ability measured by TOEIC-LC (to comprehend utterances in English) and the more complex ability assessed in the LPI situation (to comprehend and produce utterances in English).

Even though TOEIC-R score is substantially correlated with the LPI-criterion, it does not appear to be measuring criterion-related abilities that are different from those being measured the TOEIC-LC items--with which the R measure is relatively highly correlated (coefficients in the mid-.70's). This is a theoretically consistent finding: ability to comprehend spoken English is an integral aspect of the functional ability assessed in the face-to-face interview; this is not true of reading ability.

The pattern of correlational findings suggests that examinees with relatively high (low) average levels of TOEIC-assessed ability to comprehend spoken English may be expected to perform relatively well (poorly) in the interview situation, on the average, regardless of their average level of reading ability.

Results of the residual analysis reinforce this proposition.

- o When LPI was estimated from LC only, based on a general regression developed using combined data for the Japanese, French, Mexican, and Saudi samples, mean residuals for the several application samples were comparably small (none was greater than /.15/ in absolute value on the 0-5 LPI scale).

- o When LPI was estimated from a comparably derived Total-score equation (with the R component), the mean residual for the Saudi sample, was noticeably larger (/.44/) than mean residuals for the other samples (none greater than /.18/ in absolute value).

- o In the Saudi sample only, the TOEIC-LC mean was rather markedly higher than the TOEIC-R mean.

Thus, in a sample with divergent LC and R means (indicating differential levels of development of the corresponding English-language macroskills), the actual average level of criterion performance conformed to expectation based on the average level of TOEIC-assessed LC rather than expectation based on the average TOEIC-Total score (and, by inference, the much lower TOEIC-R mean). This occurred despite the fact that the within-sample TOEIC/LPI intercorrelations were very strong (all coefficients were in the mid-.80's).

High correlations between various language skills suggest that "different aspects of language tend to be learned together . . . and advancement in any aspect of language is generally accompanied by advancement in other aspects" (Carroll, 1983: 94). However, in particular subpopulations, the rate and course of development of one aspect of second-language proficiency

may differ from that in other aspects of proficiency, as illustrated by the markedly different TOEIC-LC and TOEIC-R means for the Saudi examinees.

Consistent Evidence Pointing to Functional LC/LPI Linkage

On balance, the evidence that has been reviewed suggests strongly that the ability to comprehend and produce utterances in English is to some extent "dependent," directly and functionally, upon the ability to comprehend spoken English. Accordingly, it follows logically that level of ability to use English in face-to-face conversation (indexed by LPI performance) is likely to vary relatively consistently with level of developed English-language listening comprehension (indexed by the TOEIC-LC score), across as well as within samples of ESL users/ learners from diverse TOEIC subpopulations such as those represented in the present study.

Although the relationship between reading ability and LPI performance is relatively strong, it derives indirectly from criterion related variance that is common to both the reading measure and the (functionally pertinent) listening comprehension measure. Even though performance on a measure of listening comprehension and a measure of reading ability are likely to be closely related in samples from a particular subpopulation, the corresponding English-language macroskills are necessarily equally highly developed in that subpopulation.

This suggests the "distinctness of listening and reading as traits," as concluded by Bachman and Palmer (1983) based on results of a factor study involving 10 ESL proficiency measures, 5 of reading and 5 of speaking skills (including the LPI, administered by individuals trained for the ad hoc study).[32] At the same time, measures of language macroskills are relatively strongly intercorrelated in samples of educated, ESL users/learners. This is indicated by results of the present study, results reported by Bachman and Palmer (1983), and general research findings (see Hale, 1986, for a summary of research in the TOEFL testing context; see also Pike, 1979; Oller, 1983, passim.).

Inferring LPI Performance from TOEIC-LC in the Larger TOEIC Testing Context: Conclusions

The evidence adduced in this study supports the following conclusion (thought of as a strong working hypothesis):

> the level of LPI performance associated with particular levels of performance on TOEIC Listening Comprehension is likely to be relatively consistent across diverse samples from national subpopulations characterized by differential average levels of development of TOEIC-assessed English-language listening comprehension and reading skills.

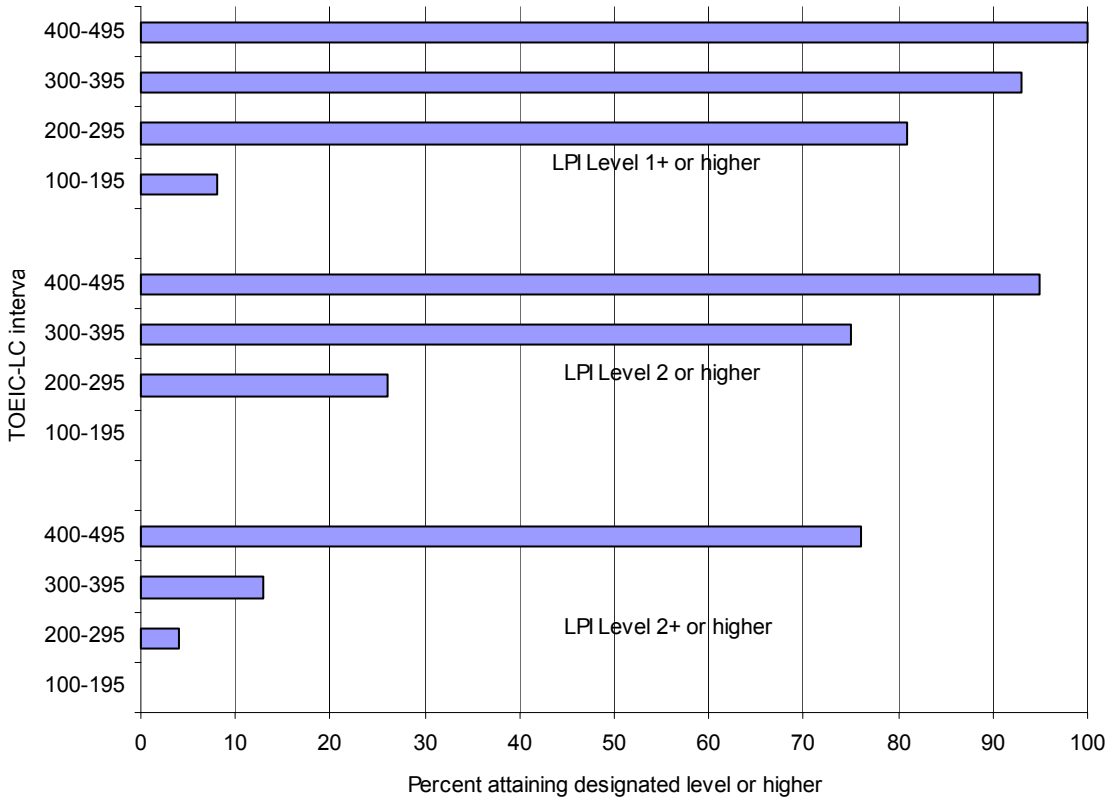
This is a necessary condition for establishing meaningful general, as opposed to "subpopulation specific," guidelines for interpretive inferences about performance on a criterion measure from predictor score(s).[33]

Based on the evidence and lines of reasoning developed above, the regression results for the combined sample of Japanese, French, Mexican, and Saudi examinees (summarized, above, in Figure 8 and Table 9), constitute guidelines that have interpretive relevance for test-users in the larger TOEIC context--certainly for general estimation purposes.

Figure 9 provides information regarding the percentage of examineees by TOEIC-LC intervals expected to earn designated LPI ratings.

o The data suggest that a substantial majority of examinees with TOEIC-LC scores of 400 or better will tend to be at LPI Level 2+ or higher, that a comparable majority of those with scores between 300 and 400 will tend to be at or above LPI Level 2, and so on.

Figure 9. Likelihood of performing at designated functional levels in LPI's in English, by score-level on TOEIC Listening Comprehension



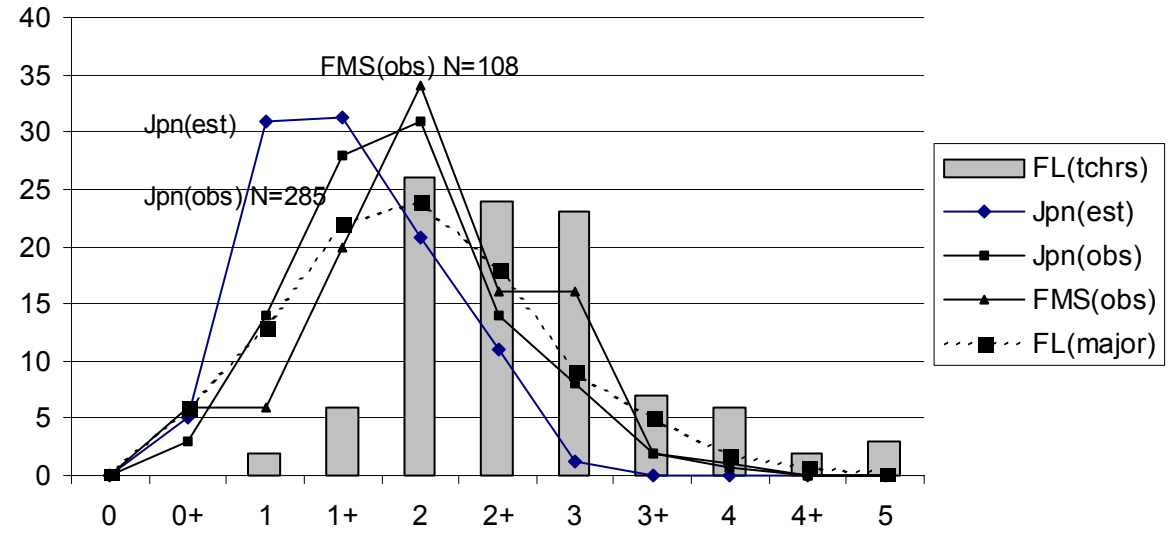
The trends highlighted in Figure 9 constitute guidelines for making inferences from examinees' TOEIC-LC performance about their probable level of LPI performance.[34] By inference, they will be able to use English in face-to-face conversation as outlined in the detailed behavioral descriptions for the corresponding oral language proficiency levels that are provided in Appendix A.

Perspective on the Distribution of LPI-Assessed Oral English Proficiency for TOEIC Examinees

Because TOEIC score data are available for general samples of Japanese examinees, it was possible, using regression equations developed in the calibration sample, to estimate the distribution of oral English proficiency according to the behaviorally defined LPI scale for the subpopulation of Japanese TOEIC examinees. Equally comprehensive TOEIC score data are not yet available for general examinee subpopulations in France, Mexico, Saudi Arabia, and most other current TOEIC subpopulations.

However, useful general normative inferences may be drawn from Figure 10, which shows relative frequency distributions (in polygon form) of LPI rating for the Japanese and the combined FMS (non-Japanese) TOEIC/LPI-calibration samples, and for a general sample of Japanese TOEIC examinees.

Figure 10. Distributions of LPI ratings in English for TOEIC samples, and of ratings in French or Spanish for samples of teachers of these languages and of college seniors specializing in these languages in the U. S.



Note. "Jpn(est)" is LPI estimated from TOEIC-LC, for SP examinees.

evaluating these distributions it is useful to recall that the FMS distribution reflects data for a sample with a TOEIC-LC mean toward the upper end of the "5-495" standard-score scale (428 for the French sample), and a sample with a comparatively low LC mean (262 for the Mexican sample). As additional score-data become available, it will be possible to make more precise estimates of score distributions for the general TOEIC testing context.

For general interpretive perspective, relative frequency distributions of LPI ratings in French or Spanish are shown for samples of secondary-school teachers of these languages in the U.S. (histogram of average ratings from Hilton et al., 1985), and U.S. college seniors specializing in the study of these languages (a relative frequency polygon adapted from estimates by Carroll, 1967).[35]

It is assumed for working purposes, that LPI ratings are generally comparable across target languages. By combining information from Figure 9 and Figure 10, it is possible to draw interpretive inferences regarding (a) the nature of the distribution of LPI-assessed ability to use English in face-to-face conversation in the general TOEIC testing context, and (b) the functional proficiency of TOEIC examinees at certain score levels relative to that of defined samples of second-language specialists--that is, language students and language teachers. For example:

o Most of the ESL users/learners likely to be tested with the TOEIC did not specialize in ESL during their educational careers. Relatively few of them are likely to earn ratings much beyond Level 3.

However, this appears to hold as well for populations made up predominately of nonnative speakers of two target languages who specialized in the study of those languages--samples of U.S. student specialists and language teachers, inference, relatively highly selected on proficiency-related variables.

o Most of the foreign language teachers demonstrated LPI-assessed oral language proficiency judged to be at or above Level 2--largely between Level 2 and Level 3, inclusive. Figure 9 shows that a majority of examinees with TOEIC Listening Comprehension scores of 300 or higher are expected to demonstrate proficiency rated at Level 2 or higher--representing the attainment of " . . . a highly usable set of skills" (ETS, 1982b: 131).

It follows that the distribution of LPI-assessed oral language proficiency for TOEIC examinees with LC scores of 300 or above is comparable to that for the focal population of "foreign language teachers."

The following comments by Lowe (1987: 45, emphasis added) are useful in light of evidence that relatively few academically-trained ESL users/learners in the TOEIC testing population are likely to be rated much above LPI Level 3. Lowe provides perspective regarding the interpretation of interview performance rated at Level 3, as well as higher levels on the LPI(ILR) scale.

The ILR scale is developmental in nature. At the summit the scale refers to the proficiency of an educated native speaker (ENS). <u>This does not imply that all natives are at Level 5</u>. ENS

> status is normally acquired through long-term familiarization (from infancy to university graduate school) with varying kinds of language and social groups over a wide number of concrete and abstract subject areas. Although most individuals at Level 5 possess a diploma, ENS status is proven by the examinee's ability to use the language. ILR experience shows that the majority of native speakers of English probably fall at Level 3. In ILR experience, the number of nonnative Level 5's is miniscule.

Comments by Carroll (1983: 102-103) about the spoken language skills of native speakers provide additional perspective on the difference between the "native-like" functional ability to exchange meaning in English that Lowe associates with LPI Level 3 and the equally "native-like" abilities associated with higher levels on the LPI scale.

> (S)tudy of the spoken language skills of native speakers may appear to be rather supererogatory, because at least at adult levels, native speakers have almost by definition acquired to a high degree the communicative skills that second language learners seek to acquire. Even young children have acquired many of these skills. Native speakers do not make the 'errors' in phonology, lexicon, and grammar that nonnatives make, even those who are fairly well advanced. If native speakers make errors in tests of 'grammar,' these tests often turn out to be tests of formalistic conventions associated with certain aspects of 'educated' speech and writing styles . . .; they represent advanced phases of language development that go beyond the normal acquisition of a second language (emphasis added).

By inference, ESL users/learners who perform at LPI Level 3 (attainable, according to Lowe, by a majority of native English speakers), have acquired native-like "communicative" skills--but not levels of advanced, "educated" English proficiency with respect to which native speakers themselves differ markedly (as evidenced, for example, by differences in performance on tests of "verbal ability" used in college admission).

Directions for Further Research on TOEIC/LPI Relationships

For the Japanese-examinee subpopulation, the strength and consistency of TOEIC/LPI relationships has been amply documented--in the present study, Woodford's (1982) validation study, and studies conducted by Japanese scholars specializing in English-language instruction and assessment (e.g., Saegusa, 1985). TOEIC/LPI relationships are similarly strong in initial samples from three emergent TOEIC subpopulations.

One can conclude, as a strong working hypothesis, that the pattern of TOEIC/LPI (predictor/criterion) relationships observed in these samples will be consistent across similarly selected samples of ESL users/learners in major national subpopulations in the larger TOEIC testing context. There is theoretical as well as evidentiary support for so concluding. However, evidence regarding TOEIC/LPI relationships in samples from other developed and developing TOEIC subpopulations is needed to permit a comprehensive empirical evaluation of this working hypothesis. TOEIC/LPI data-sets for samples from additional countries probably will be

generated naturalistically, as the data-sets employed in the present study were generated, during the course of operational assessments in TOEIC-use settings.

As additional TOEIC/LPI data-sets become available from diverse settings, it will be possible to obtain empirical answers to the two questions that appear to be most pertinent:

1. Is the pattern of TOEIC/LPI concurrent correlations consistent with that observed in the samples under consideration in the present study?

2. Is the average level of criterion performance consistent with expectation based on the average level of performance on the TOEIC (as specified by guidelines developed in the general TOEIC/LPI-calibration sample available for the present study)?

Potential Usefulness of Self-Ratings of Oral English Proficiency

There is reason to believe that self-ratings of oral English proficiency may be a useful surrogate for actual interview ratings in research designed to assess consistency of patterns of relationships between TOEIC scores and level of oral English proficiency across diverse national subpopulations and to identify nontest variables that may contribute to the prediction of LPI performance, after controlling for TOEIC scores.

For example, there is evidence suggesting that "adult learners can assess their speaking and listening skills in much the same way as do their teachers" (Ingram, 1985: 268). Hilton et al. (1985) reported relatively high correlations between self-ratings of speaking proficiency (in Spanish and French) and the corresponding LPI ratings--correlations were r = .66 and r = .69 in samples of French and Spanish teachers in the United States.

Results of a self-assessment substudy. More direct evidence bearing on the usefulness of self-ratings of oral English proficiency as a surrogate criterion (for purposes of research) in the TOEIC testing context is provided by results of a special self-assessment substudy based on data collected by TOEIC/ETS staff in the sample of French examinees (see Table 7 and related discussion, above). Self-ratings of oral English proficiency were obtained using a rating scale with LPI-parallel level-descriptions (see Appendix B, Exhibit B.2).

An analysis was made of interrelationships among TOEIC scores, self-ratings, and LPI ratings. Selected findings are shown in Table 10.

For present purposes, the most pertinent aspect of these findings is that the general level and pattern of relationships between TOEIC scores and self-rated oral English proficiency was quite similar to that between TOEIC scores and the actual LPI-criterion measure (other aspects of these findings are discussed in Appendix B). Assuming that oral English proficiency is more closely related to TOEIC-LC than to TOEIC-R, the results obtained using self-ratings as the criterion and those obtained using actual LPI ratings (the surrogate criterion) lead to the same conclusion.

Table 10

Selected Findings of the Self-Assessment Substudy in the French Sample (N = 56)

Variable	Correlation with Self rating	Correlation with Interview rating	Mean	SD
TOEIC-LC	.640	.616	428	74
TOEIC-R	.501	.583	389	48
TOEIC-Total	.628	.646	817	113
Self rating	-	.643	2.77	.93
LPI rating	.643	-	2.30	.64
Pred LPI.lc*	.640	.616	2.46	.39

*LPI estimated from TOEIC-LC using the combined-sample regression equation (see Table 8, above, and related discussion).

Self-ratings of oral English proficiency (and other aspects of proficiency such as writing or reading) are obtainable on a routine basis as part of the regular test administration process, along with the responses of examinees to pertinent background questions (sex, age, educational level, extent of use of English on-the-job, job categories, type of employer, time spent in an English-speaking environment, and so on).[36] Self-ratings could be used as a surrogate criterion in studies designed to identify demographic, experiential, or other variables that contribute to the prediction of self-assessed oral English proficiency, after controlling for TOEIC scores. The results of such studies should provide a basis for assessing and formulating working hypotheses regarding non-test variables that may contribute to estimation of LPI performance after controlling for performance on the TOEIC.[37]

Further exploration of the usefulness of self-assessments of oral English proficiency, using a rating scale with LPI-parallel level descriptions, is warranted (as is attention to the development of procedures for collecting data on potentially relevant personal, demographic, and experience variables in all major testing contexts--along lines now well-established in the Japanese testing context).

Other Directions for Future TOEIC Research

This study was not designed to obtain evidence regarding the relationship of TOEIC scores to directly assessed measures of English-language reading or writing skills in samples of TOEIC examinees. Woodford (1982) provided evidence that TOEIC scores were closely related to direct measures of both of these skills, as well as to LPI performance. However, Woodford did not rate the samples of reading and writing ability according to behaviorally defined levels that paralleled those of the LPI oral language proficiency scale.

Do Equal TOEIC-LC and TOEIC-R Scores Reflect "Comparable Levels of Proficiency?"

In his study of the attainments of foreign language majors in the U.S., based on estimated functional levels of oral language proficiency and reading ability, Carroll (1967) concluded that foreign language majors were generally less advanced in listening and speaking skills than in reading and writing skills.

> Typically, the 'regular' cases had mean scores in Listening and Speaking that correspond to FSI ratings of S-2 or S-2+, i.e., in the range of 'limited working proficiency.' In Reading and Writing, however, the tested students tended to have mean scores that correspond approximately to an FSI rating of R-3 . . ." (p. 199).

Moreover, based on evidence introduced in Section II (see especially Figures 2a, 2b, & 2c, above), educated ESL users/learners in the TOEFL testing context tend to be considerably more advanced, on the average, in reading ability than in the ability to comprehend spoken English or, by inference, to use English conversationally.

Judging the foregoing, the academically trained ESL users/learners in the TOEIC testing context may tend to be more native-like, on the average, in their functional ability to read (and possibly to write) English, than they are in their ability to comprehend and produce utterances in English.

To the extent that this is true, standard scores on TOEIC-LC and TOEIC-R (scores representing equal deviations from standardization-sample raw-score means on the respective measures) may represent different levels of functional ability. It is important to obtain evidence bearing on this issue.[38]

The question of differences in level of skill-development could be addressed by obtaining evidence regarding the comparative performance of native-English-speaking counterparts of, say, Japanese test-takers, on the LC and R sections of the TOEIC--for example, in cooperative studies designed to obtain TOEIC-score distributions for native-English-speaking employees and Japanese employees in comparable positions (sales, engineering, and so on) with particular companies.

This question might also be addressed by using procedures for the direct assessment of reading (and writing) skills. Such procedures do not appear to have been widely used in operational testing settings. Carroll's (1967) use of the LPI-parallel procedure for rating reading proficiency may represent a unique application in the context of a large-scale assessment of functional levels of second-language skills in general populations of second-language users/ learners.

Assessment of developmental levels of reading and writing skills entails problems of behavior sampling that are not present to the same extent in assessing LPI performance. Detailed examination of problems associated with the application of LPI-parallel procedures for the direct assessment of reading and writing skills is beyond the scope of this paper. However, it is

pertinent to note that the controlled conversational interview is particularly useful as a basis for eliciting and rating second-language proficiency. This is so, in part, because it is an interactive assessment procedure that allows the interviewer to elicit and evaluate behavior in any area (e.g., functioning, register) deemed to be relevant for establishing an examinee's functional command of a target language. On the other hand, in collecting samples of writing ability, for example, it is inherently more difficult to obtain a correspondingly representative sampling of pertinent behavior.[39]

It is possible that self-ratings of reading and writing ability according to a schedule with LPI-parallel level-descriptions might prove useful for research purposes in the TOEIC testing context. In the case of writing ability, graded samples of general business correspondence might be used as part of the self-assessment process.

Need to Translate General Interpretive Guidelines into Context-Specific Interpretive Guidelines

The chain of interpretive inference that has been validated in this study is definitionally limited. The evidence explicitly links TOEIC scores, directly, to one very clearly defined and generally important aspect of developed ability to use English as educated native speakers may be expected to use the language, namely, the ability to use English in face-to-face conversation as reflected in performance in Language Proficiency Interviews conducted under controlled conditions by trained interviewers/raters.

Setting Local Interpretive Guidelines

Ultimately the information conveyed by TOEIC scores and LPI ratings needs to be linked (formally-statistically and/or clinically-intuitively) to defined criteria of the ability of ESL users/learners to use English in the workplace. This would entail evaluative judgments regarding the adequacy or relative adequacy of the performance of employees in specific ESL-dependent-positions--that is, positions in which successful job performance is dependent to some extent upon ability to use English.

Questions at issue in the workplace generally have to do with establishing the implications of test performance for ESL-proficiency-related selection, placement, training, and job-classification decisions. Such decisions must be made within constraints imposed by a finite pool of employees or prospective employees with a particular joint distribution of English language skills, the amount of time and resources available for training designed to improve skill levels in the pool, and other practical considerations.

The process of developing meaningful local (context-specific) interpretive guidelines needs to be guided by

1. a realistic assessment of the extent to which incumbents in ESL-dependent positions are meeting the assignments associated with those positions in a manner that is considered satisfactory by general company standards,

2. the assumption that different jobs require different levels and patterns of proficiency in English, and

3. the premise that decisions regarding test-based minimum proficiency requirements should take into account the actual score distributions of employees whose overall on-the-job performance is judged, by usual company standards, to be at least minimally satisfactory--minimum proficiency requirements for getting the job done are likely to vary depending upon the job.

In connection with the last point, it is important to recognize that the characterization of a particular level on the LPI scale as representing the attainment of "minimum working proficiency" or "minimum professional proficiency" should not be thought of as a general guideline for making workplace decisions about levels of proficiency "required" for "successful" performance in particular ESL-dependent positions.

Form-and-Substance versus Substance in Communication

The LPI scale evaluates linguistic behavior in terms of native speaker norms (expectations), not specifically in terms of level of functional "communicative competence" or "achievement of mutual intelligibility," to use Savignon's (1986) terminology.

> (There is a distinction between) adoption of native speaker norms--writing or speaking like a native speaker . . . --and the achievement of mutual intelligibility--communicating with native speakers" (Savingnon, 1986: p. 21); . . . (It is important to determine) the extent to which deviations of various kinds from native speaker norms interfere with mutual intelligibility. Psycholinguistic studies provide ample evidence that utterances may be interpretable without being "natural," i.e., native-like, and that semantically deviant utterances are more likely to be misinterpreted than are grammatically deviant utterances (pp. 22-23, emphasis added in all instances).

The findings of this study (see especially Figure 10 and related discussion, above) suggest that only a small proportion of TOEIC examinees (with quite atypical English-language backgrounds) are likely to be able to write or speak like a native speaker of English--that is, to meet "native speaker levels" on the LPI scale. Acceptance of the distinction made by Savignon implies pragmatic realism, not a "lowering of standards."

In essence, LPI-scaled proficiency levels, like TOEIC scores, need to be validated against criteria of ability to accomplish workplace assignments that are contingent upon demonstrated ability to establish and maintain on-line communicative interaction at a level of "mutual intelligibility" that is sufficient for accomplishing the (business-related, or other) purposes of the interaction.

The extent to which individuals below LPI Level 2 (or with TOEIC-LC scores below 300) are able to meet the communicative requirements of positions involving different levels of interaction with native-English speakers is an important empirical question.[40]

SECTION V: GENERAL CONCLUDING OBSERVATIONS

The Language Proficiency Interview procedure has very strong face validity as a measure of general ability to use English in face-to-face conversation. The linguistic demands imposed by participation in the interview situation are in many ways very similar to the linguistic demands associated with exchanging meaning on-the-job, in situations calling for communicative interaction in English. It thus constitutes a relevant general criterion for establishing the representational value of the TOEIC--that is, an expectancy-set based on knowledge of test-criterion relationships as to how well an individual is likely to be able to use English.

It is evident that interpretive dividends have been realized in the TOEIC testing context by the use of a regression-based criterion-referencing model in which performance in Language Proficiency Interviews was used as a general context-independent criterion. This was expected, a priori, because scores on the LPI criterion have direct representational value, and the regression model, by definition, can be expected to indicate the extent to which TOEIC scores share the criterion measure's representational value.

Knowledge of TOEIC/LPI relationships represents a clear interpretive advance because it permits test users to make statistically valid inferences from employees' TOEIC scores about their levels of developed oral English proficiency (as illustrated in Figure 9, above, for example). It also provides better-informed perspective regarding the level and range of oral English proficiency that academically trained ESL users/learners in the TOEIC testing context can be expected to exhibit (as illustrated in Figure 10, above).

These interpretive dividends--that obviously will be shared by TOEIC users and others interested in second-language assessment--have accrued from the TOEIC program's long-term investment in the development and maintenance of a strong direct assessment program to complement its major program of norm-referenced testing. The TOEIC experience in providing comprehensive assessment services in Japan and elsewhere in the world indicates clearly that it is feasible for a large-scale ESL proficiency testing program to develop and maintain a strong operational capability for the direct assessment of oral English proficiency. This is achieved by offering in strategic locations the type of education, training, and periodic "recalibration" needed to facilitate the development of a cadre of program-related, resident ESL professionals highly skilled in the use of the LPI procedure.

The TOEIC direct assessment program has contributed novel evidence regarding the probable level and range of functional ability to use English in face-face-conversation for a potentially very large population of ESL users/learners. This evidence clarifies the effective range of developed oral English proficiency being assessed by the TOEIC.

Finally, the results of this study attest to the elemental clarity of Carroll's (1967) insight that the interpretive power inherent in behaviorally scaled direct assessments could be harnessed --by empirical linkage rules established in samples from defined populations--to psychometrically more efficient norm-referenced measures of language macroskills, and thus be extended to the populations involved.[41]

References

Adams, M. L. (1978). Measuring foreign language speaking proficiency: Study of agreement among raters. In J. L. D. Clark (Ed.). (1978b). Direct testing of speaking proficiency: Theory and application (129-150). Princeton, NJ: Educational Testing Service.

American Association of Collegiate Registrars and Admissions Officers (1971). AACRAO-AID Participant Selection and Placement Study. Report to the Office of International Training, Agency for International Development, U. S. Department of State. Washington, D. C.: Author.

Alderson, C. J., Krahnke, K. J., & Stansfield, C. W. (Eds.). (1987). Reviews of English Language Proficiency Tests. Washington, DC: Teachers of English to Speakers of Other Languages.

Angelis, P. J., Swinton, S. S., & Cowell, W. R. (1979). The performance of non-native speakers of English on TOEFL and verbal aptitude tests (TOEFL Research Reports, Report 3). Princeton, NJ: Educational Testing Service.

Angell, A. G., Gallagher, A. M., & Schneider, L. M. (1988). Test analysis: Test of English as International Communication (ETS SR-88-26). Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Sharon, A. T. (1971). Comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants to U.S. colleges. TESOL Quarterly, 5, 129-136.

Bachman, L. F., & Palmer, A. S. (1983). The construct validity of the FSI Oral Interview. In J. W. Oller (Ed). Issues in language testing research (154-169). Rowley, Massachusetts: Newbury House.

Bailey, K. M., Dale, T. L, & Clifford, R. T. (Eds.). (1987). Language Testing Research: Selected Papers from the 1986 Colloquium. Monterey, CA: Defense Language Institute.

Bejar, I. (1985). A preliminary study of raters for the Test of Spoken English (TOEFL Research Reports, Report 18, and ETS-RR-83-32). Princeton, NJ: Educational Testing Service.

Bragger, J. D. (1985). The development of oral proficiency. Northeast Conference Reports, 41-75.

Breland, H. (1983). The direct assessment of writing skill: A measurement review (College Board Report No. 83-6, and ETS RR-83-32). NY: College Entrance Examination Board.

Breland, H. (1977). A study of college English placement and the Test of Standard Written English (College Board Report RDR-76-77, No.4, and ETS Project Report, PR-77-1).

Brumfit, C. (Ed.). (1982). English for International Communication, N.Y.: Pergamon.

Campbell, R. (1986). Discussion of the Savignon and Candlin papers, and the response by Larsen-Freeman (p. 61). In Toward communicative competence testing: Proceedings of the Second TOEFL Invitational Conference (TOEFL Research Reports, No. 21). Princeton, NJ: Educational Testing Service.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English (TOEFL Research Reports, Report 19, GRE Board Research Report 83-R2, and ETS RR-85-21). Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1967). The foreign language attainments of language majors in the senior year: A survey conducted in U..S. colleges and universities (Final Report, Contract OE-4-14-048). Cambridge, MA: Graduate School of Education, Harvard University. (Also available as Eric Document 013343).

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller (Ed). Issues in language testing research (80-107). Rowley, Massachusetts: Newbury House.

Cartier, F.L. (1975). Discussion of Wilds, C. (1975: 42). In R. L. Jones, & B. Spolsky (Eds.), Testing language proficiency (29-44). Arlington, VA: Center for Applied Linguistics.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones, & B. Spolsky (Eds.), Testing language proficiency (10-28).Arlington, VA: Center for Applied Linguistics.

Clark, J. L. D. (1978a). Psychometric considerations in language testing. In B. Spolsky (Ed). Approaches to language testing (15-30). Arlington, VA: Center for Applied Linguistics.

Clark, J. L. D. (Ed.). (1978b). Direct testing of speaking proficiency: Theory and application. Princeton, NJ: Educational Testing Service.

Clark, J. L. D. (1978c). Interview testing research at Educational Testing Service. In J. L. D. Clark (Ed.), Direct testing of speaking proficiency: Theory and application (211-228). Princeton, NJ: Educational Testing Service.

Clark, J. L. D. (1987). A study of the comparability of speaking proficiency interview ratings across three government language training agencies. In K. M. Bailey, T. L. Dale, & R. T. Clifford (Eds.), Language Testing Research: Selected Papers from the 1986 Colloquium (132-179). Monterey, CA: Defense Language Institute.

Clark, J. L. D., & Swinton, S. S. (1979). An exploration of speaking proficiency measures in the TOEFL context (TOEFL Research Report s, Report 4). Princeton, NJ: Educational Testing Service.

Clark, J. L. D., & Swinton, S. S. (1980). The Test of Spoken English as a measure of communicative ability in English-medium instructional settings (TOEFL Research Reports, Report 7). Princeton, NJ.: Educational Testing Service.

Cowell, W. R. (1980). Test analysis for Test of English for International Communication (ETS Statistical Report SR-80-13). Princeton, NJ: Educational Testing Service.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). TOEFL from a communicative viewpoint on language proficiency: A working paper (TOEFL Research Reports, Report 17). Princeton, NJ: Educational Testing Service.

Educational Testing Service (1981). TOEFL test and score manual. Princeton, NJ: Author.

Educational Testing Service (1982a). Test of English for International Communication: Bulletin of Information. Princeton, NJ: Author.

Educational Testing Service (1982b). ETS oral proficiency testing manual. Princeton, NJ: Author.

Educational Testing Service (1983). TOEFL test and score manual. Princeton, NJ: Author.

Educational Testing Service (1985a). TOEFL test and score manual. Princeton, NJ: Author.

Educational Testing Service (1985b). Test of English for International Communication: Bulletin of information. Princeton, NJ: Author.

Educational Testing Service (1985c). Test of Spoken English: Manual for users. Princeton, NJ: Author.

Educational Testing Service (1986a). Guide for TOEIC users. Princeton, NJ: Author.

Educational Testing Service (1986b). TOEIC International Corporate Program. Princeton, NJ: Author.

Educational Testing Service (1988). Full range of English as a second language assessment programs under way at ETS, ETS Developments, XXXII, 5-8.

Hale, G. (1986). An overview of research related to TOEFL. In C. W. Stansfield (Ed.), Toward communicative competence testing: Proceedings of the second TOEFL invitational conference (TOEFL Research Report, Report 21) [10-16]. Princeton, NJ: Educational Testing Service.

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). Summaries of studies involving the Test of English as a Foreign Language, 1963-1982 (TOEFL Research Reports No. 16). Princeton, NJ: Educational Testing Service.

Hilton, T. L., Grandy, J., Kline, R. G, & Liskin-Gasparro, J. E. (1985). The oral language proficiency of teachers in the United States in the 1980's--An empirical study. Princeton, NJ.: Educational Testing Service.

Hyltenstam, K., & Pienemann, M. (Eds.). (1985). Modelling and assessing second language acquisition. San Diego, CA: College-Hill Press.

Ingram, D. E. (1985). Assessing proficiency: An overview of some aspects of testing. In K. Hyltenstam, & M. Pienemann (Eds.), Modelling and assessing second language acquisition (215-276). San Diego, CA: College-Hill Press.

Ito, Akira (1987). Personal communication.

Jones, R. L. (1978). Interview techniques and scoring criteria at the higher proficiency levels. In J. L. D. Clark (Ed.). (1978b). Direct testing of speaking proficiency: Theory and application (89-102). Princeton, NJ: Educational Testing Service.

Jones, R. L., & Spolsky, B. (1975). (Eds.) Testing language proficiency. Arlington, VA: Center for Applied Linguistics.

Lado, R. (1978). Scope and limitations of interview-based language testing: Are we asking too much of the interview? In J. L. D. Clark (Ed.). (1978b). Direct testing of speaking proficiency: Theory and application (113-128). Princeton, NJ: Educational Testing Service.

Livingston, S. A. (1978). Setting standards of speaking proficiency. In J. L. D. Clark (Ed.). (1978b). Direct testing of speaking proficiency: Theory and application (255-270). Princeton, NJ: Educational Testing Service.

Lowe, P. (1987). Interagency Language Roundtable Oral Proficiency Interview. In C. J. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), Reviews of English Language Proficiency Tests (43-47). Washington, DC: Teachers of English to Speakers of Other Languages.

Oller, J. W. (1983). (Ed.). Issues in language testing research. Rowley, MA: Newbury House.

Perkins, K. (1987). Test of English for International Communication. In C. J. Alderson, K. J. Krahnke, & C. English Language Proficiency Tests (81-83). Washington, DC: Teachers of English to Speakers of Other Languages.

Pike, L. W. (1979). An evaluation of alternative item formats for testing English as a foreign language (TOEFL Research Report 2, and ETS RR-79-6). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Stansfield, C. W. (1985). Testing the oral English proficiency of foreign nursing graduates. The ESP Journal, 4, 21-35.

Quinn, T. J., & McNamara, T. F. (1987). Australian Second Language Proficiency Ratings. In C. J. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), Reviews of English Language Proficiency Tests (7-9). Washington, DC: Teachers of English to Speakers of Other Languages.

Reilley, V. (1988). Personal communication.

Reschke, C. (1978). Adaptation of the FSI Interview Scale for secondary schools and colleges. In J. L. D. Clark (Ed.). (1978b). Direct testing of speaking proficiency: Theory and application (75-88). Princeton, NJ: Educational Testing Service.

Saegusa, Y. (1983). Japanese college students' reading proficiency in English, Musashino English and American Literature, 16, 99-117. (Tokyo, Japan: Musashino Women's University).

Saegusa, Y. (1985). Prediction of English Proficiency Progress, Musashino English and American Literature, 18, 165-185. (Tokyo, Japan: Musashino Women's University).

Saegusa, Y. (1989). Japanese company workers' English proficiency, WASEDA Studies in Human Sciences, 2, 1-12.

Savignon, S. (1986). The meaning of communicative competence in relation to the TOEFL program. In C. W. Stansfield (Ed). Toward communicative competence testing: Proceedings of the Second TOEFL Invitational Conference (TOEFL Research Reports, No. 21). Princeton, NJ: Educational Testing Service.

Sollenberger, H. E. (1978). Development and current use of the FSI Oral Interview Test. In J. L. D. Clark (Ed.). (1978b). Direct testing of speaking proficiency: Theory and application (1-12). Princeton, NJ: Educational Testing Service.

Stansfield, C. W. (1986). (Ed.) Toward communicative competence testing: Proceedings of the second TOEFL invitational conference (TOEFL Research Report 21). Princeton, NJ: Educational Testing Service.

Thompson, R. T. (1985). Testing, standards, and the curriculum: Through the backdoor. In K. R. Jankowsky (Ed.) Scientific and humanistic dimensions of language, 149-156. Amsterdam: John Benjamin Publishing Co.

Thorndike, R. L. (1949). Personnel Selection: Test and Measurement Techniques. New York: John Wiley.

Wilds, C. P. (1975). The Oral Interview Test. In R. L. Jones, & B. Spolsky (Eds.), Testing language proficiency (29-44). Arlington, VA: Center for Applied Linguistics.

Wilson, K. M. (1986). The relationship of scores based on GRE General Test item types to undergraduate grades: An exploratory study for selected subgroups (GRE Board Professional Report GREB No. 83-19P & ETS Research Report 86-37). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1984). The relationship of GRE General Test item-type part scores to undergraduate grades (GRE Board Professional Report GREB No. 81-22P & ETS Research Report 84-38). Princeton, NJ: Educational Testing Service.

Woodford, P. E. (1982). The Test of English for International Communication (TOEIC). In C. Brumfit (Ed.), English for International Communication, 61-72: N.Y.: Pergamon Press.

APPENDICES

APPENDIX A Levels of Oral English Proficiency in the FSI/ILR Scale (see, e.g., Duran, Canale, Penfield, & Stansfield, 1985)

Preface

The following proficiency level descriptions characterize *spoken language use*. Each of the six "base levels" (coded 00, 10, 20, 30, 40, and 50) implies control of any previous "base level's" functions and accuracy. The "plus level" designation (coded 06, 16, 26, etc.) will be assigned when proficiency substantially exceeds one base skill level and does not fully meet the criteria for the next "base level." The "plus level" descriptions are therefore supplementary to the "base level" descriptions.

A skill level is assigned to a person through an authorized language examination. Examiners assign a level on a variety of performance criteria exemplified in the descriptive statements. Therefore, the examples given here illustrate, but do not exhaustively describe, either the skills a person may possess or situations in which he/she may function effectively. Statements describing accuracy refer to typical stages in the development of competence in the most commonly taught languages in formal training programs. In other languages, emerging competence parallels these characterizations, but often with different details.

Unless otherwise specified, the term "native speaker" refers to native speakers of a standard dialect.

"Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education. However, in cultures where formal higher education is common, the language-use abilities of persons who have had such education is considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language.

Speaking 0 (No Proficiency)

Unable to function in the spoken language. Oral production is limited to occasional isolated words. Has essentially no communicative ability. (Has been coded S-0 in some nonautomated applications.)

Speaking 0+ (Memorized Proficiency)

Able to satisfy immediate needs using rehearsed utterances. Shows little real autonomy of expression, flexibility, or spontaneity. Can ask questions or make statements with reasonable accuracy only with memorized utterances or formulae. Attempts at creating speech are usually unsuccessful.

Examples: The individual's vocabulary is usually limited to areas of immediate survival needs. Most utterances are telegraphic; that is, functors (linking words, markers, and the like) are omitted, confused, or distorted. An individual can usually differentiate most significant sounds when produced in isolation, but, when combined in words or groups `of words`, errors may be frequent. Even with repetition, communication is severely limited even with people used to dealing with foreigners. Stress, intonation, tone, etc. are usually quite faulty. (Has been coded S-0+ in some nonautomated applications.)

Speaking 1 (Elementary Proficiency)

Able to satisfy minimum courtesy requirements and maintain very simple face-to-face conversations on familiar topics. A native speaker must often use slowed speech, repetition, paraphrase, or a combination of these to be understood by this individual. Similarly, the native speaker must strain and employ real-world knowledge to understand even simple statements/questions from this individual. This speaker has a functional, but limited proficiency. Misunderstandings are frequent, but the individual is able to ask for help and to verify comprehension of native speech in face-to-face interaction. The individual is unable to produce continuous discourse except with rehearsed material.

Examples: Structural accuracy is likely to be random or severely limited. Time concepts are vague. Vocabulary is inaccurate, and its range is very narrow. The individual often speaks with great difficulty. By repeating, such speakers can make themselves understood to native speakers who are in regular contact with foreigners, but there is little precision in the information conveyed. Needs, experience, or training may vary greatly from individual to individual; for example, speakers at this level may have encountered quite different vocabulary areas. However, the individual can typically satisfy predictable, simple, personal, and accommodation needs; can generally meet courtesy, introduction, and identification requirements; exchange greetings; elicit and provide, for example, predictable and skeletal biographical information. He/she might give information about business hours, explain routine procedures in a limited way, and state in a simple manner what actions will be taken. He/she is able to formulate some questions even in languages with complicated question constructions. Almost every utterance may be characterized by structural errors and errors in basic grammatical relations. Vocabulary is extremely limited and characteristically does not include modifiers. Pronunciation, stress, and intonation are generally poor, often heavily influenced by another language. Use of structure and vocabulary is highly imprecise. (Has been coded S-1 in some nonautomated applications.)

Speaking 1+ (Elementary Proficiency, Plus)

Can initiate and maintain predictable face-to-face conversations and satisfy limited social demands. He/she may, however, have little understanding of the social conventions of conversation. The interlocutor is generally required to strain and employ real-world knowledge to understand even some simple speech. The speaker at this level may hesitate and may have to change subjects due to lack of language resources. Range and control of the language are limited. Speech largely consists of a series of short, discrete utterances.

Examples: The in dividual is able to satisfy most travel and accommodation needs and a limited range of social demands beyond exchange of skeletal biographic information. Speaking ability may extend beyond immediate survival needs. Accuracy in basic grammatical relations is evident, although not consistent. May exhibit the more common forms of verb tenses, for example, but may make frequent errors in formation and selection. While some structures are established, errors occur in more complex patterns. The individual typically cannot sustain coherent structures in longer utterances or unfamiliar situations. Ability to describe and give precise information is limited. Person, space, and time references are often used incorrectly. Pronunciation is understandable to natives used to dealing with foreigners. Can combine most significant sounds with reasonable comprehensibility, but has difficulty in producing certain sounds in certain positions or in certain combinations. Speech will usually be labored. Frequently has to repeat utterances to be understood by the general public. (Has been coded S-1+ in some nonautomated applications.)

Speaking 2 (Limited Working Proficiency)

Able to satisfy routine social demands and limited work requirements. Can handle routine work-related interactions that are limited in scope. In more complex and sophisticated work-related tasks, language usage generally disturbs the native speaker. Can handle with confidence, but not with facility, most normal, high-frequency social conversational situations including extensive, but casual conversations about current events, as well as work, family, and autobiographical information. The individual can get the gist of most everyday conversations but has some difficulty understanding native speakers in situations that require specialized or sophisticated knowledge. The individual's utterances are minimally cohesive. Linguistic structure is usually not very elaborate and not thoroughly controlled; errors are frequent. Vocabulary use is appropriate for high-frequency utterances, but unusual or imprecise elsewhere.

Examples: While these interactions will vary widely from individual to individual, the individual can typically ask and answer predictable questions in the workplace and give straightforward instructions to subordinates. Additionally, the individual can participate in personal and accommodation-type interactions with elaboration and facility; that is, can give and understand complicated, detailed, and extensive directions and make non-routine changes in travel and accommodation arrangements. Simple structures and basic grammatical relations are typically controlled; however, there are areas of weakness. In the commonly taught languages, these may be simple markings such as plurals, articles, linking words, and negatives or more complex structures such as tense/aspect usage, case morphology, passive constructions, word order, and embedding. (Has been coded S-2 in some nonautomated applications.)

Speaking 2+ (Limited Working Proficiency, Plus)

Able to satisfy most work requirements with language usage that is often, but not always, accept-able and effective. The individual shows considerable ability to communicate effectively on topics relating to particular interests and special fields of competence. Often shows a high degree of fluency and ease of speech, yet when under tension or pressure, the ability to use the language effectively may deteriorate. Comprehension of normal native speech is typically nearly complete. The individual may miss cultural and local references and may require a native speaker to adjust to his/her limitations in some ways. Native speakers often perceive the individual's speech to contain awkward or inaccurate phrasing of ideas, mistaken time, space, and person references, or to be in some way inappropriate, if not strictly incorrect.

Examples: Typically the individual can participate in most social, formal, and informal interactions; but limitations either in range of contexts, types of tasks, or level of accuracy hinder effectiveness. The individual may be ill at ease with the use of the language either in social interaction or in speaking at length in professional contexts. He/she is generally strong in either structural precision or vocabulary, but not in both. Weakness or unevenness in one of the foregoing, or in pronunciation, occasionally results in miscommunication. Normally controls, but cannot always easily produce, general vocabulary. Discourse is often incohesive. (Has been coded S-2+ in some nonautomated applications.)

Speaking 3 (General Professional Proficiency)

Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Nevertheless, the individual's limitations generally restrict the professional contexts of language use to matters of shared knowledge and/or international convention. Discourse is cohesive. The individual uses the language acceptably, but with some noticeable imperfections; yet, errors virtually never interfere with understanding and rarely disturb the native speaker. The individual can effectively combine structure and vocabulary to convey his/her meaning accurately. The individual speaks readily and fills pauses suitably.

In face-to-face conversation with natives speaking the standard dialect at a normal rate of speech, comprehension is quite complete. Although cultural references, proverbs, and the implications of nuances and idiom may not be fully understood, the individual can easily repair the conversation. Pronunciation may be obviously foreign. Individual sounds are accurate; but stress, intonation, and pitch control may be faulty.

Examples: Can typically discuss particular interests and special fields of competence with reasonable ease. Can use the language as part of normal professional duties such as answering objections, clarifying points, justifying decisions, understanding the essence of challenges, stating and defending policy, conducting meetings, delivering briefings, or other extended and elaborate informative monologues. Can reliably elicit information and informed opinion from native speakers. Structural inaccuracy is rarely the major cause of misunderstanding. Use of structural devices is flexible and elaborate. Without searching for words or phrases, the individual uses the language clearly and relatively naturally to elaborate concepts freely and make ideas easily understandable to native speakers. Errors occur in low-frequency and highly complex structures. (Has been coded S-3 in some nonautomated applications.)

Speaking 3+ (General Professional Proficiency, Plus)

Is often able to use the language to satisfy professional needs in a wide range of sophisticated and demanding tasks.

Examples: Despite obvious strengths, may exhibit some hesitancy, uncertainty, and effort, or errors which limit range of language-use tasks that can be reliably performed. Typically there is particular strength in fluency and one or more, but not all, of the following: breadth of lexicon, including low- and medium-frequency items, especially socio-linguistic/cultural references and nuances of close synonyms; structural precision, with sophisticated features that are readily, accurately, and appropriately controlled (such as complex modification and embedding in Indo-European languages); discourse competence in a wide range of contexts and tasks, often matching a native speaker's strategic and organizational abilities and expectations. Occasional patterned errors occur in low-frequency and highly complex structures. (Has been coded S-3+ in some nonautomated applications.)

Speaking 4 (Advanced Professional Proficiency)

Able to use the language fluently and accurately on all levels normally pertinent to professional needs. The individual's language usage and ability to function are fully successful. Organizes discourse well, using appropriate rhetorical speech devices, native cultural references, and understanding. Language ability only rarely hinders him/her in performing any task requiring language; yet, the individual would seldom be perceived as a native. Speaks effortlessly and smoothly and is able to use the language with a high degree of effectiveness, reliability, and precision for all representational purposes within the range of personal and professional experience and scope of responsibilities. Can serve as an informal interpreter in a range of unpredictable circumstances. Can perform extensive, sophisticated language tasks, encompassing most matters of interest to well-educated native speakers, including tasks which do not bear directly on a professional specialty.

Examples: Can discuss in detail concepts which are fundamentally different from those of the target culture and make those concepts clear and accessible to the native speaker. Similarly, the individual can understand the details and ramifications of concepts that are culturally or conceptually different from his/her own. Can set the tone of interpersonal official, semi-official, and non-professional verbal exchanges with a representative range of native speakers (in a range of varied audiences, purposes, tasks, and settings). Can play an effective role among native speakers in such contexts as conferences, lectures, and debates on matters of disagreement. Can advocate a position at length, both formally and in chance

encounters, using sophisticated verbal strategies. Understands and reliably produces shifts of both subject matter and tone. Can understand native speakers of the standard and other major dialects in essentially any face-to-face interaction. (Has been coded S-4 in some nonautomated applications.)

Speaking 4+ (Advanced Professional Proficiency, Plus)

Speaking proficiency is regularly superior in all respects, usually equivalent to that of a well-educated, highly articulate native speaker. Language ability does not impede the performance of any language-use task. However, the individual would not necessarily be perceived as culturally native.

Examples: The individual organizes discourse well, employing functional rhetorical speech devices, native cultural references, and understanding. Effectively applies a native speaker's social and circumstantial knowledge. However, cannot sustain that performance under all circumstances. While the individual has a wide range and control of structure, an occasional non-native slip may occur. The individual has a sophisticated control of vocabulary and phrasing that is rarely imprecise, yet there are occasional weaknesses in idioms, colloquialisms, pronunciation, cultural reference, or there may be an occasional failure to interact in a totally native manner. (Has been coded S-4+ in some nonautomated applications.)

Speaking 5 (Functionally Native Proficiency)

Speaking proficiency is functionally equivalent to that of a highly articulate well-educated native speaker and reflects the cultural standards of the country where the language is natively spoken. The individual uses the language with complete flexibility and intuition, so that speech on all levels is fully accepted by well-educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references. Pronunciation is typically consistent with that of well-educated native speakers of a non-stigmatized dialect. (Has been coded S-5 in some nonautomated applications.)

End Appendix A

APPENDIX B

Reliability and Self-Assessment Substudies

The analyses reported in this appendix pertain to two areas that were not directly at issue in this study, namely, (a) the reliability of LPI ratings employed as the criterion measure, and (b) the usefulness of self-ratings of "oral English proficiency" according to an LPI-parallel rating schedule, as a research-surrogate for actual LPI ratings.

A Reliability Substudy

Observed correlations in the mid-.70s between LPI ratings and TOEIC scores provide strong indirect evidence of relatively high levels of inter-rater reliability. In connection with the assessment conducted in Mexico (see Table 7 and related discussion) it was possible to obtain some direct evidence regarding inter-rater reliability.

The TOEIC-ETS interviewer-rater who conducted the interviews rated them on the spot and recorded the ratings. At a later time, a second TOEIC-ETS staff member rated the audiotaped protocols and recorded the ratings independently. A consensus rating was then obtained and employed as the criterion measure. Using the separately recorded ratings it was possible to obtain direct evidence of inter-rater reliability. The correlation between ratings (N = 42) was .90. There were differences in ratings in 17 cases; no difference exceeded .5. Rater means were quite close: 1.71 versus 1.66.

This result clearly represents a high level of agreement both in rank-order and level for the pair of raters involved. By inference from the consistently strong TOEIC/LPI correlations obtained in the Japanese samples, the ratings of LPI performance that were generated by the cadre of TOEIC\ETS-trained interviewer-raters in Japan (at the Institute for International Studies and Training, and elsewhere) appear to have had a similarly high degree of reliability (or more generally, reproducibility). These ratings are possibly comparable to levels reported by Adams (1978) for ratings generated within the Foreign Service Institute, under "same roof" conditions.

It is relevant to recall at this juncture that all the interviewers-raters involved were trained initially in the LPI technique, and periodically "recalibrated," through participation in TOEIC-ETS training sessions conducted by the same individual, namely, Steven A. Stupak. Certain critical aspects of "being under the same roof" appear to be represented in these circumstances.

In evaluating the foregoing, it is of interest to note that in the Hilton et al. (1985) study, average inter-rater reliabilities of .71 and .73 were reported in circumstances in which "under the same roof" conditions were <u>not</u> present. Ratings were made by "scattered-site" teams of interviewers-raters recruited for an ad hoc study of second-language oral proficiency in samples of U.S. secondary-school teachers of Spanish and French, respectively.

A Self-Assessment Substudy

A distinguishing feature of the Hilton et al. (1985) study was the fact that self-ratings of speaking proficiency (in Spanish or French) were referenced to LPI-scale in a calibration substudy (through equipercentile equating). The correlation between the self-rating and the obtained LPI rating was .66 in the French calibration sample and .69 in the corresponding Spanish sample. No test data were employed.

The study findings indicated that by referencing self-assessments of oral language proficiency to LPI ratings, useful estimates of functional ability to use the target language as defined by the LPI descriptors could be derived. However, the descriptions of behavior associated with each of six levels of the self-rating schedule employed did not conform strictly to the LPI scale. The extent to which the mean self-rating differed from the mean LPI rating was not a matter of concern. A copy of the self-rating schedule employed for French teachers is shown as Exhibit B.1 (the schedule for Spanish teachers was identical except for the language designations).

Given the generic interpretive contribution of the behaviorally linked LPI scale, it is important to assess degree of agreement be- tween LPI ratings and self-assessments that are based on a schedule of behaviorally defined levels paralleling the official LPI scale, as to basic content and number of levels. Such a scale was developed and used as part of the ad hoc ESL assessment, conducted by Steven A. Stupak and other TOEIC-ETS staff, involving employees in ESL-essential positions in the Paris office of a TOEIC corporate client (see Table 7 in the text, and related discussion). As might be expected, the second-language reading load of the all-category rating schedule (written in English) was relatively heavy (see Exhibit B.2).

Selected findings of analyses involving the self-report data are provided in Table B.1. Several points are noteworthy:

1. The pattern of relationships between TOEIC scores and self-assessed speaking proficiency was very similar to the pattern of TOEIC/LPI relationships. For example, LC correlated more closely than R with each criterion, the relative size of the two coefficients was about the same for each criterion, and so on. This is consistent with the correlation of .64 between the self-ratings and the interview ratings.

2. TOEIC-Total and LC scores were about as highly correlated with the LPI criterion as was the self-assessment.

3. From a "scaling" perspective (concerned with the fit between LPI mean and self-assessment mean), the average self-rating was higher by about .5 than the actual LPI mean.

4. The actual LPI mean in the French sample was estimated with equal accuracy from TOEIC-LC and from TOEIC-Total, using combined-sample calibration equations (for perspective, see text, Table 8 and related discussion).

Exhibit B.1

Self-Rating Schedule Used by Hilton, et al. (1985) in a National (U.S.) Sample of Teachers of French

If you DO NOT speak any FRENCH, please check here and then skip to Question 32.

29. This question asks you to judge your own level of speaking ability in French. Please read each one of the six paragraphs below and decide which paragraph best describes your ability to speak and to understand spoken French. Please be as honest and as accurate as possible. Below paragraph 6, in the space provided, write the number preceding only the one paragraph that best describes your speaking ability in French. If you believe that your speaking ability in French is between levels, choose the lower level (e.g., the lower numbered paragraph).

1. My speech in French is limited to a few words and I have great difficulty understanding French, even when it is spoken very slowly. I cannot really communicate any information in the language.

2. I can ask and answer questions about very familiar subjects and can understand simple questions and statements if they are spoken slowly, and sometimes repeated. My vocabulary is limited to basic needs (food, asking directions, greeting people, and so forth). I make many grammatical mistakes but can usually be understood by French speakers who deal with foreigners. I can order food in a restaurant, get a room in a hotel, ask directions on the street, and introduce myself to people.

3. I can talk with native speakers of French about myself and my family, my job, studies, or hobbies. I can recount a story and describe an event. I can understand most conversations in French except when the speech is very fast. My grammar is fairly good but I make mistakes with complicated constructions. If I do not know the word for a particular thought or object, I can usually describe it by using other, easier words.

4. I can talk about professional topics with ease, and am able to state and support my opinions. I can understand almost everything spoken by native French users. My vocabulary is good enough so that I usually know most or all of the words for what I want to say. My grammar is good and any mistakes I make are usually with the more complicated constructions. My pronunciation is good but may not be completely native.

5. I can talk fluently and accurately about almost any subject with which I am familiar, including professional, abstract, or controversial topics. I can always understand native French speakers, even when they are speaking quickly and using sophisticated or colloquial expressions. My vocabulary is very extensive, and I make only a very few grammatical errors. My pronunciation is very good but may not be completely native.

6. My speech is exactly like that of an educated native speaker of French.

Paragraph # _____ best describes my speaking ability in French.

Exhibit B.2

Self-Rating Schedule Developed by TOEIC/ETS Staff

Identification Information: Last Name(s):_____ First Name: _____
Location: _____ Identification No.:_____
TOEIC Test Date : Mo ---------- Day ---------- Year: 19 ----
DO NOT WRITE IN THIS BOX L ------- R ------- T -----
LPI/Qrre: No. -------
INSTRUCTIONS

Please read down the following list quickly until you find a description that corresponds somewhat to your language level.

Then carefully read both up and down from that level, until you find the description that corresponds most accurately to your assessment of your language level. If you are uncertain as to how to reply, please ask your English teacher to help you.

Thank you for your cooperation.

1. MY ENGLISH IS LIMITED TO MEMORIZED WORDS AND PHRASES. I ASK QUESTIONS ONLY WHEN I THINK I KNOW THE ANSWER OR WHEN THE ANSWER IS YES OR NO. I AM NOT ABLE TO USE GRAMMAR.

I am able to satisfy immediate survival needs using memorized material. There are long pauses in my speech and I must rely on single words and phrases for all communication, except for occasional uses of one or two verbs. I can ask questions and make statements with reasonable accuracy only by using short memorized material. My vocabulary is limited to survival needs. I can understand sounds when they are isolated, but when they are in words or groups of words, I have a very difficult time understanding them. When I talk with people in English, even if they are used to speaking with learners. I have a very difficult time making myself understood.

2. I AM ABLE TO USE LIMITED ENGLISH GRAMMAR. I MUST TRANSLATE EVERYTHING I SAY BUT I CAN MAKE SENTENCES. I KNOW ENOUGH ENGLISH TO SURVIVE IN THE LANGUAGE, IF PUT IN ENGLISH-SPEAKING SITUATION.

I am able to survive in English, to order a meal and get a room in a hotel. I can ask and answer simple questions, ask for directions, respond to simple statements, and maintain very simple face-to-face conversation. I am able to ask questions using limited grammar, but usually with much inaccuracy. Almost all of my statements contain fractured syntax and other grammatical errors. If I repeat what I say, I can make myself understood to people who have regular contact with learners of English. If I think about it, I am able to create sentences, using simple verbs, nouns, pronouns, etc.

3. I AM ABLE TO SPEAK ENGLISH WITHOUT TRANSLATING EVERY SENTENCE. I KNOW SOME PAST TENSE AND FUTURE TENSE VERBS BUT I AM UNCERTAIN IN THEIR USE. I FEEL COMFORTABLE IN INFORMAL SOCIAL SITUATIONS, IF NOT ASKED TO SPEAK VERY MUCH.

I am able to speak English as described above, as well as satisfy limited social demands. I am able to produce language more spontaneously, but it is still difficult for me to speak. I am able to use some verbs in the past and future tenses. Because my vocabulary is limited, I often have to think about words. I am unable to control grammar in longer sentences and/or unfamiliar situations. I have only limited ability to describe and cannot give precise information. In conversation, I speak in short sentences, thinking about almost every one before speaking. I have difficulty producing certain sounds, but generally my speech is understood. While I am not comfortable speaking English for any length of time, I think that with some practice I could become a confident speaker.

3. I CAN DISCUSS MY WORK, HOME LIFE, CURRENT EVENTS, HOBBIES, LIKES AND DISLIKES IN ENGLISH. I CAN UTTER SENTENCES IN SERIES, WITHOUT HAVING TO PAUSE, AND CAN NARRATE AND DESCRIBE IN PAST, PRESENT AND FUTURE. MY CONTROL OF BASIC GRAMMAR, INCLUDING PAST TENSE VERBS, IS GOOD.

I am able to speak English as described above, but I am also able to discuss my work in English. I can discuss matters that are of a concrete nature with some confidence. I am comfortable in social situations, casual conversations, discussing current events, work, family, and my likes and dislikes. I can understand most conversations on non-technical subjects, if they require no specialized information. I can give detailed directions on how to get from one place to another. I am able to narrate and describe in the past, present, and future, and I can string sentences together in conversation with no difficulty. I do not have a thorough or confident control of all grammar, but I am able to use simple language quite accurately. While my vocabulary is not very precise, it is broad enough to discuss a wide variety of topics, most of which are familiar to me. In simple conversations, I rarely have to mentally translate what I am going to say.

5. I "KNOW" ALMOST ALL OF THE ENGLISH GRAMMAR, BUT I AM NOT YET ABLE TO USE IT. I USE SIMPLE GRAMMAR WITH NO PROBLEM BUT EXPERIENCE DIFFICULTY WITH MORE ADVANCED GRAMMAR. MY COMPREHENSION IS ALMOST 100%. WHEN EXPRESSING AN OPINION, I HAVE TO STOP AND THINK ABOUT HOW I AM GOING TO EXPRESS IT.

I am able to speak English as described above, but my vocabulary is adequate to discuss a wide range of topics, many of which are unfamiliar to me. I have quite good control of verbs in the past and future, and am able to use other less frequent forms, including conditionals. I still make mistakes with regard to grammar and vocabulary that I consider simple, but those errors are more the result of carelessness than anything else. I have been exposed to all of the important grammar points, but I cannot always recall them as I speak. I find it difficult to discuss concepts, thoughts, opinions, or hypothetical situations in English, as I am required to both think

and speak at the same time. When I offer an opinion, etc., I have to stop and think about how I will express myself. My accent is quite good and I am very fluent, except when I must stop to think about what I am going to say.

6. I AM ABLE TO SAY NEARLY EVERYTHING I WANT TO SAY IN ENGLISH, ALTHOUGH MY VOCABULARY IS NOT ALWAYS PRECISE. I CAN EXPRESS OPINIONS, DISCUSS CONCEPTS, AND HYPOTHESIZE. I AM ABLE TO USE ALL GRAMMATICAL CONSTRUCTIONS, BUT MAKE SOME MISTAKES.

I am able to participate effectively in most formal and informal conversations on practical, social, and professional topics. I can discuss particular interests and special fields of competence with reasonable ease. I understand nearly everything I hear in normal, everyday speech. My vocabulary is broad enough that if I do not know a word, it does not prevent me from expressing myself. One way or another, I am able to express myself an every topic on which I wish to express myself and I am never driven to silence because of lack of control of grammar or vocabulary limitations. Grammatical errors in my speech are still fairly common, but they never prevent me from being understood. I am able to discuss any topic that is presented, on which I would be able to comment in my native language.

7. I CAN USE ALL VERB FORMS CONSISTENTLY AND ACCURATELY, BUT SOMETIMES MAKE MISTAKES. I HAVE MINOR PROBLEMS WITH CERTAIN GRAMMAR POINTS, SUCH AS ARTICLES AND PREPOSITIONS. I CAN FUNCTION EFFECTIVELY IN UNFAMILIAR SITUATIONS.

I am able to speak as described above, but the error rate in my speech is quite low. My control of grammar is such that I can use all verb forms consistently and accurately, although there are still certain patterns of error in my speech. I am able to speak English with a structural accuracy and a breadth of vocabulary sufficient to extensively discuss my professional needs. I understand and use a great many idiomatic expressions, but in that regard my speech is still foreign. I sometimes have difficulty with certain verb forms, articles, and prepositions. I would not be taken for a native speaker, but I am able to respond appropriately in unfamiliar or surprise situations.

8. I MAKE ONLY OCCASIONAL RANDOM ERRORS IN ENGLISH. MY VOCABULARY IS BROAD AND PRECISE. MY FORM OF ADDRESS IS ALWAYS APPROPRIATE WITH REGARD TO LEVEL OF LANGUAGE, WHETHER WITH CHILDREN, PERSONAL FRIENDS, GOVT. OFFICIALS, EDUCATORS, ETC.

I am able to use English fluently and accurately on all levels required by my profession. I am able to address all parties appropriately, whether in formal or informal situations. I can pattern my speech appropriately to express compliments, condolences, surprise, disappointment, and affection. I can participate in any conversation within the range of my personal and professional experience with a high degree of fluency and precision of vocabulary. Errors in pronunciation and grammar are quite rare. There are no patterns of error in my speech, and while I may make errors in speaking, they are only random, sporadic errors and do not constitute definable weaknesses in my speech. I can handle informal interpretation both from and into English,

working with my native language. Curiously enough, my speech is often much more correct than that of many native speakers, although I realize, especially when I sit down to write, that there are still areas in which I need more study and practice. I am able to hide most of my language problem areas sufficiently well that native speakers do not know what they are and in casual conversation usually cannot detect them. I still need to work on specialized vocabulary in a number of areas in which I am well versed in my native language. Most of my problems are with vocabulary, regional accents, low frequency idiomatic expressions, and trying to understand uneducated native speakers.

9. I RARELY MAKE MISTAKES IN MY SPEECH. IT DIFFERS FROM THAT OF THE NATIVE SPEAKER BECAUSE OF AN OCCASIONAL MISTAKE IN GRAMMAR, VOCABULARY, ACCENT, OR IDIOMATIC EXPRESSION. SOME PEOPLE MAY SOMETIMES MISTAKE ME FOR BEING A NATIVE SPEAKER.

My speech is sometimes equivalent to that of an educated native speaker of English, but I am not able to sustain that level for extended periods of time. My use of vocabulary, colloquialisms, and cultural references and not always entirely appropriate, although I rarely make mistakes in those areas. I feel nearly as comfortable in English as I do in my native language, and native speakers never feel they have to modify their speech, regardless of what they are discussing, for me to understand it. My English is nearly flawless.

10. MY SPEECH IS INDISTINGUISHABLE FROM THAT OF THE EDUCATED NATIVE SPEAKER.

My speech is equivalent to that of an educated native speaker of English in every regard--pronunciation, fluency, breadth and precision of vocabulary, grammar, comprehension of slurred, regional, colloquial, and other non-standard speech, and cultural referents.

Note. For interpretive perspective only (not included in the operational form of the scale):
The numbered descriptions in this self-assessment schedule correspond to ILR levels, paraphrased for self-assessment purposes, as follows:

1 = Level 0 plus, 2 = Level 1, 3 = Level 1 plus, 4 = Level 2,
5 = Level 2 plus, 6 = Level 3, 7 = Level 3 plus, 8 = Level 4,
9 = Level 4 plus, 10 = Level 5.

Table B.1

Correlation of TOEIC Scores with LPI Rating and Self Rating, Respectively

Variable	Correlation with LPI rating	Correlation with Self rating	Mean	SD
TOEIC-LC	.640	.616	428	74
TOEIC-R	.501	.583	389	48
TOEIC-Total	.628	.646	817	113
Self rating	--	.643	2.77	.93
LPI rating	.643	--	2.30	.64
Pred LPI.tot*	.628	.646	2.45	.33
Pred LPI.lc**	.640	.616	2.46	.39

* LPI rating estimated from TOEIC Total, using the regression equation developed using data for the combined sample (N=393) of Japanese, French, Mexican, and Saudi examinees.

**LPI rating estimated from TOEIC Total, using the regression equation developed using data for the combined sample (N=393) of Japanese, French, Mexican, and Saudi examinees.

These findings indicate that further exploration of self-ratings based on LPI-parallel behavioral descriptions is warranted. They suggest, as do the Hilton et al. findings, that a substantial, quite useful degree of agreement in rank order between self-ratings and actual LPI ratings can be expected. Of course, questions naturally arise regarding the extent to which adult, educated ESL learners/users are likely to be able to place themselves on the FSI scale, after referring to FSI-parallel behavioral descriptions, as they are placed by expert judgment based on actual interviews.

It is important to consider the impact of the heavy reading load (in English, or other target language) imposed by a schedule such as that shown in Exhibit B.1. One way to reduce that load would be to develop a schedule using the native language of the ESL users/learners involved; another would be to explore the usefulness of abbreviated versions of the descriptors.

Development of additional evidence regarding TOEIC/LPI relationships in major TOEIC-use contexts clearly is quite important from both theoretical and practical perspectives. Self-ratings might well be obtained on a routine basis, along with responses to background questions (sex, age, educational level, extent of use of English on-the-job, years of language study, time spent in English-speaking environments, type of work, and so on). Questions on the specific nature of the English-language demands of positions would be particularly useful (e.g., primarily reading English-language technical journals; frequent, direct interaction with native-speakers of English).

Findings of studies designed to identify demographic, experiential, or other variables that contribute to the prediction of self-assessments, after controlling for TOEIC scores, should provide a basis for useful working hypotheses regarding the concomitants of LPI behavior.

APPENDIX C

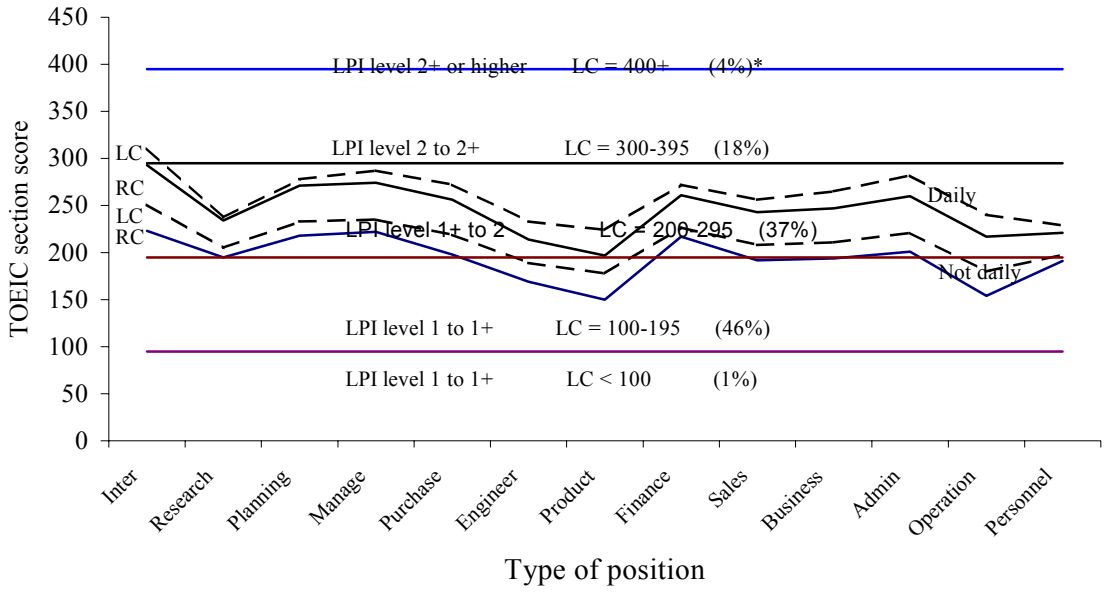
Illustrative Data from the TOEIC Testing Context in Japan

Saegusa (1989) reports that TOEIC-using companies have set target goals in terms of TOEIC Total scores for certain positions. For example, the goal for newly employed college graduates is TOEIC-Total scores averaging about 450 (in the 400 to 500 range); for engineers, the target is an average of 600 (scores in the 500 to 650 range), and for international jobs, the goal is an average of 650 (scores in the 600 to 730 range).

However, he points out that only 13 percent of all examinees meet the minimum target scores set for sales or engineering, and only 9 percent meet the targets for international positions. Moreover, the average score reported for 14,292 engineers (in a total sample of 67,792 examinees) was only 396; for over 10,000 examinees in sales jobs the average was 430, and so on (see Exhibit C.1).

Figure C.1 (based on data from Exhibit C.1) shows mean scores for TOEIC-LC and R, respectively, for Japanese TOEIC examinees classified by type of position and by frequency of use of English (daily versus other). Positions are ordered (from left to right) by percentage of incumbents reporting daily use of English (high = 89 percent, low = 20 percent). It is assumed for purposes of discussion that individuals reporting daily use of English are in ESL-dependent positions with their respective companies.

Figure C.1. Mean TOEIC LC and RC scores by type of position and frequency of use of English (daily vs. other): Japanese IP examinees (data from Saegusa, 1989)



** Estimates from present study (total sample)*

Exhibit C.1

TOEIC Performance of Japanese Institutional Program Examinees by Position and Frequency of Use of English (Saegusa, 1989)

QUESTION 1 : *Do you use English every day either at work or outside work?*

	ANSWERED YES					ANSWERED NO				
JOBS	N	%	L	R	T	N	%	L	R	T
MANAGEMENT	89	48.6	287	274	561	94	51.4	235	222	456
ADMINISTRATION	335	29.0	282	260	542	821	71.0	221	201	422
FINANCE	492	33.8	272	261	533	962	66.2	227	217	443
PERSONNEL	424	20.3	229	221	449	1663	79.7	198	191	389
PLANNING	822	52.8	278	271	549	735	47.2	233	218	451
INTERNATIONAL	1718	89.2	311	293	604	209	10.8	251	223	474
PURCHASING	200	43.0	272	256	528	265	57.0	219	198	417
BUSINESS	1337	29.0	265	247	511	3268	71.0	211	194	404
SALES	3125	29.9	256	243	499	7315	70.1	208	192	401
ENGINEERING	6043	42.3	233	214	447	8249	57.7	189	169	358
PRODUCTION	587	35.2	224	197	421	1082	64.8	178	150	328
RESEARCH	3714	63.7	237	234	471	2120	36.3	205	195	401
OPERATIONS	136	25.1	240	217	457	405	74.9	180	154	334
NO ANSWER	4936	22.9	223	202	425	16646	77.1	194	177	371
TOTAL	23958	35.3	245	230	475	43834	64.7	199	182	382

QUESTION 2 : *Have you lived in an English-speaking country, using English as a means of communication, for an accumulated period of six months or over?*

	ANSWERED YES					ANSWERED NO				
JOBS	N	%	L	R	T	N	%	L	R	T
MANAGEMENT	24	13.0	345	302	648	160	87.0	247	238	486
ADMINISTRATION	54	4.7	376	316	692	1098	95.3	232	213	445
FINANCE	105	7.2	376	333	708	1347	92.8	231	224	455
PERSONNEL	67	3.2	359	300	659	2020	96.8	200	193	393
PLANNING	104	6.7	365	328	693	1451	93.3	249	240	489
INTERNATIONAL	477	24.8	365	329	694	1444	75.2	284	271	556
PURCHASING	33	7.1	362	309	671	431	92.9	232	217	449
BUSINESS	216	4.7	370	314	684	4384	95.3	219	204	423
SALES	525	5.0	357	302	659	9885	95.0	216	202	418
ENGINEERING	607	4.2	322	284	606	13678	95.8	203	184	387
PRODUCTION	71	4.3	299	252	551	1597	95.7	190	162	352
RESEARCH	188	3.2	358	321	679	5638	96.8	221	217	438
OPERATIONS	19	3.5	348	286	634	522	96.5	189	166	355
NO ANSWER	729	3.3	378	308	686	21327	96.7	195	178	373
TOTAL	3219	4.7	357	307	664	64982	95.3	209	194	402

Saegusa (1989: 2) reports that: "Japanese companies are not satisfied with the employees' current level of English proficiency." The source of that dissatisfaction appears to be the discrepancy between targeted test levels and actual test levels.

Only 16 percent of all examinees have TOEIC-LC or –R scores above 300, or Total scores of 600 or above (Total LC + R). Many interesting questions are raised by the data summarized in Figure C.1.

Are the Goals "Realistic"?

According to the general rationale for setting local interpretive guidelines, outlined briefly in the body of this report, it is important to determine what proportion of employees in various ESL-dependent positions (EDP) are "getting the job done" to the satisfaction of their employers. It is also relevant to ask, for example, whether engineers (in a highly technical specialization) and incumbents in sales positions need the same level of general oral English proficiency to be successful in their positions. Unless, based on the record, a significant percentage of the incumbent work force in EDPs in the respective areas is "not getting the job done", some reassessment of target goals would appear to be called for.

In this connection, one can only agree with Saegusa's (1989: p. 1) observation that: "So far Japanese industries seem to have been doing very well in international business".

NOTES TO TEXT

Section I

1. The terms "English as a second language" (ESL) and "English as a foreign language" (EFL) are used interchangeably in this report, in a generic sense, to indicate that the study is concerned with acquired proficiency in English in samples of nonnative speakers--individuals for whom English is a foreign or second language--regardless of the purpose for which they acquired, are using, or expect to use English (e.g., whether they use it on a casual basis for personal and social reasons, or daily for travel, work, or study in an English-speaking environment). The focal population of test takers is composed of highly educated, adult ESL users/learners whose patterns of ESL acquisition include a core of academic exposure to the study of English as a foreign language.

2. For example, although the TOEFL (ETS, 1985a) is very widely used to screen ESL applicants for admission to U.S. colleges and universities, and the users are advised to make followup studies designed to link TOEFL score levels to ESL-communication-related performance criteria, no examples of such institutional studies were found in a comprehensive summary of research involving the TOEFL between 1963 and 1982 (Hale, Stansfield, & Duran, 1984). See Ingram (1985: 237-239) for a commentary on the problem of conducting research designed to establish the functional implications of scores on indirect, norm-referenced tests.

3. To supplement a search of the literature, the writer queried Professor Carroll, by telephone, as to whether he knew of any subsequent study of this kind. Professor Carroll said that he knew of none, and attributed the apparent lack of replication to the costs and the administrative and logistical difficulties involved in obtaining the direct assessments.

Section II

4. Problems of this nature are inherent in validation research involving context-specific performance criteria--that is, they are not peculiar to studies involving context-specific criteria of ability to use a target language to accomplish defined language-essential tasks. In assessing the predictive validity of norm-referenced college admission tests, for example, studies relating test scores to first-year grade point average (GPA)--a "context-specific" index of academic accomplishment--need to be conducted in each setting in which the test scores are used.

5. For a rare example of a validity study involving a contextspecific criterion, see Clark and Swinton (1980). The relatively scaled criterion variables employed in the study were students' ratings of the oral English communication skills of nonnative-English speaking teaching assistants. See Livingston (1978) and Powers and Stansfield (1985) for examples of what may be termed "quasi-pragmatic" validation studies (judgments of "adequacy for particular purposes" based on recorded LPI samples and "speaking" samples respectively, not on actual observation of behavior in the real-life contexts under consideration).

6. A regression-based model can be expanded by including variables other than indirect test scores that may be considered relevant for estimating criterion behavior (e.g., age, sex, educational level, years of formal study of the target language, and so on). Generally speaking,

all variables found to account for a significant proportion of the criterion variance could reasonably be retained in a fully expanded regression model.

7. For historical perspective see Adams, 1978; Sollenberger, 1978. For additional perspective and broad evaluation of procedure see, for example Clark, 1978b, passim; Jones and Spolsky, 1975, passim. For a detailed operational manual for oral proficiency testing involving the basic FSI model and adaptations developed by ETS and the American Council on the Teaching of Foreign Languages, Inc. (ACTFL) for use in academic language-learning settings see ETS (1982b); see also Bragger (1985), and Thompson (1985). For a critical review of the LPI procedure see Lowe (1987). A comparable model has also been applied in assessing second-language reading and writing ability. Since rated behavior in conversational interviews constitutes the criterion employed in the present study, attention is focused primarily on the LPI procedure for assessing "oral language proficiency." The more recently developed Australian Second Language Proficiency Ratings model (ASLPR) also provides behaviorally anchored ratings for basic macroskills (Ingram, 1985; Quinn & McNamara, 1987).

8. Adding .5 (rather than a value nearer the next higher level) appears to have been used within the Foreign Service Institute (see Adams, 1978: 146-148), as well as in other applied or research contexts (see, for example, Carroll, 1967; Reschke, 1978; Clark and Swinton, 1979; Hilton, et al., 1985). This may seem anomalous because, as noted by Reschke (1978: 83), a plus rating is assigned ". . . only to a performance that substantially exceeds the minimum requirements for a given level but does not meet all the minimum requirements for the next higher level." Exceptions to the "add .5" practice may occur, however. Woodford (1982), for example, used .7 for intermediate ratings in a study concerned with assessing the concurrent relationships between scores on the TOEIC and LPI ratings. Accordingly, in evaluating relatively rare published reports that include mean FSI ratings, it is important to consider the possibility that some other arbitrarily selected numerical conversion (e.g., ".7" or ".8") may have been made. All such numerical conversions refer to a "plus" value and not to a specific model-related "quantification" of within-level proficiency.

9. These references are to a selected population of second-language users/learners: individuals employed by the U.S. government. Little direct empirical evidence is available regarding the distribution of LPI-defined oral language proficiency in various populations of second-language users/learners. However, second-language learners/ users in academic settings tend to be concentrated at or below Level 1. The LPI scale was expanded at the low end, through a collaborative effort by the Foreign Service Institute, the American Council of Teachers of Foreign Languages (ACTFL), and Educational Testing Service. For historical perspective and a detailed description of levels in the resulting ACTFL/ETS scale, see ETS, 1982b; also Bragger, 1985.

10. Cartier's remarks were made in the context of questions raised by discussants (Clark, Lado, Oller, Spolsky, et al.) of Wilds' (1975) paper as to whether " . . . the interview is valid for more than per-formance in an interview." For example, "To what extent have there been studies of the accuracy of judgments made on the basis of FSI [LPI] interviews? To what extent is there follow-up work, to what extent is there feed-back, when examinees go out into a real-world situation?" According to Wilds such studies apparently had not been made within the LPI-

assessment context (Wilds, 1975: 38-46, passim). Few such studies appear to have been made outside government circles. See Lado (1978) for a discussion of questions regarding the reliability and the validity of the "oral interview test," and the relative merits of indirect and direct assessment procedures from the perspective of a proponent of indirect assessment techniques. See Ingram (1985) for a similar discussion from the perspective of a proponent of direct tests.

11. It is unlikely that there has been another comparable analysis of the particular set of variables for which intercorrelations are shown in Exhibit B. However, comparably high levels of intercorrelation have been reported for generally similar sets of indirect and direct measures in samples composed primarily of educated, adult ESL users/learners. For example, see Hale (1986), for a review of such relationships in the TOEFL testing context; see also Pike (1979), Oller (1983: passim).

12. When the "equal standard deviations" equivalencies were calculated, " . . . in a number of cases, [the MLA scores calculated as corresponding to behaviorally anchored ratings were found to exceed] the maximum possible scores. This finding would suggest that the MLA tests in those cases do not have a high enough 'ceiling,' that is, that they do not have the capacity to discriminate among the upper levels of FSI ratings or indeed among the upper levels of language competence (near-native and native language ability)" (Carroll, 1967: 15). This suggests the elemental contribution of ratings of LPI performance and Reading skills on conceptually comparable developmental scales from the point of view of defining operationally the level and range of second-language proficiency in given populations and for defining the ranges of developed abilities that are being assessed by test items selected so as to be of "average difficulty" for the populations involved.

13. In connection with the findings for the French sample, Carroll noted (1967: 46) that " . . . the rating standards of the persons who judged the French speaking test results may have been for some reason unusually severe." Scoring of the MLA Speaking Test, like the FSI direct assessment procedure, calls for subjective judgment.

14. Of course, it is possible to infer differential relative levels of development of these skills in population subsamples with significant-ly different means on measures of particular macroskills.

15. See Clark (1978c) and Clark & Swinton (1979) for evidence suggesting that inter-rater correlations in the .7 - .8 range may be typical for interviewers/raters representing a relatively wide range of experience, and in ad hoc rating contexts (as opposed to “same roof” contexts, for which higher levels of inter-rater reliability have been reported [Adams, 1978]).

Section III

16. TOEIC examinees in Japan are largely university-educated and they share a basic core of exposure to curriculum-embedded English-language instruction. According to information supplied by the TOEIC Steering Committee in Japan (Ito, 1987, personal communication), the typical Japanese university graduate has had approximately 1,000 formal class-room hours of instruction in English as a foreign language, spread over a span of some 8 years, beginning with

middle school, distributed as follows:

Middle school: 3 hrs/week, by 35 weeks, by 3 years = 315 hrs.
High school : 5 hrs/week, by 35 weeks, by 3 years = 525 hrs.
University : 3 hrs/week, by 30 weeks, by 2 years = 180 hrs.

This represents a substantial core of required study of English. The level of developed functional ability to use English, conversationally and otherwise, in the TOEIC examinee population thus may be thought of as reflecting outcomes of formal English-language instruction during the period of formal schooling completed by these employ-ed adults, plus general post-graduate experiential change, plus effects that may be associated with selection into the TOEIC population**.**

17. These professionals are encouraged to participate periodically in the workshop series. Logically, periodic participation in training or recalibration sessions conducted by the same experienced interviewer/ rater is conducive to the maintenance of consistent rating standards. In essence, certain aspects of "being under the same roof" are present in these circumstances (see Adams, 1978). Trainees are tested with a set of specially prepared taped interviews, which they rate independently, submitting their ratings for "recalibration." Theoretically, the periodic "recalibration" of raters is particularly important in contexts in which the raters are more or less isolated from an environment dominated by educated native speakers of the language being rated.

18. Unless otherwise indicated, the descriptions in this section are based on information supplied by Akira Ito (a member of the staff of the TOEIC Steering Committee in Japan), under whose general supervision the TOEIC sample was selected and tested, and by Vincent Reilley, Director of the IIST English Department and responsible for the ongoing program of ESL training at the Institute. Some members of the IIST English staff were included among the interviewer/raters who were involved in collecting data for the TOEIC sample.

19. In a multiple discriminant analysis involving four groups and three test (or other) measures, three statistically independent discriminating functions (linear combinations of the measures) are derived. In this case, none of the three discriminant functions proved to be statistically significant ($p > .50$ for each function).

20. The anomalous negative regression coefficient for the reading score in the IIST-86 sample may be explained in terms of sampling fluctuation. However, it is noteworthy that in this particular sample, the listening score alone is actually more closely related to the LPI criterion than is the Total score (with the Reading component).

21. 100-point class-intervals were used for Total and 50-point intervals were used for Listening, defined in such a way that the mid-points corresponded to the selected Total and LC scores shown in Figure 4a and Figure 4b.

22. Reading ability is assessed to some extent by the TOEIC-LC measure. The point has been made (Perkins, 1987: 82) as follows: "The TOEIC is an integrative test in the sense that it

engages different modes and language components. For example, in the Listening Comprehension section, the subject reads the options in English, choosing the correct answer based on what was heard on tape."

23. The SP data were for a spaced (every-third-case) sample (N = 3,558) files maintained at ETS (Princeton). An LPI rating was estimated from the TOEIC-Total score equation for each member of the sample. Individuals were assigned to LPI levels according to class intervals of estimated LPI ratings as follows: <.75 = 0+, .75 1.24 = 1, 1.25 1.74 = 1+, and so on. Choice of the Total-score equation rather than the LC-equation for estimation purposes was arbitrary. Score data for individual IP examinees were not available for this study.

24. Examinees with extremely low, TOEIC scores were not represented in the calibration sample. By inference, very few of these examinees are likely to exhibit LPI performance ratable above Level 0+. The LPI scale does not provide discrimination among individuals at very low levels of proficiency. In order to discriminate adequately among examinees at very low levels of developed English-language proficiency, a modified version of the LPI procedure--the ACTFL/ETS version, for example (ETS, 1982b)--would be appropriate. Moreover, for such examinees an easier TOEIC-like norm-referenced measure would undoubtedly provide more efficient measurement.

25. Data for the non-Japanese samples were obtained in the ad hoc asessments, conducted by TOEIC-ETS staff members trained in the LPI procedure.

26. A regression-based calibration equation minimizes the discrepancies between observed LPI-criterion performance and performance estimated from a particular TOEIC score or weighted composite of scores (that is, LC and R), in a particular calibration sample. A given equation reflects not only the strength of association between the variables involved in the calibration sample, but also the means (and standard deviations) on both the LPI criterion and the TOEIC variable(s) in that sample. Consistency of estimation across diverse samples, from a particular regression equation, is thus dependent not only on consistency of within-sample correlations, but also on consistency of fit across samples between average level of criterion (LPI) performance and average level of TOEIC performance.

27. The estimation equations were as follows:

FMS.LC: (.004416*L) + .48861; J.LC: (.006062*L) .+ 049348;
Comb.LC: (.005332*L) + .18138.
FMS.LC+R: (.004110*L) + (.000036*R) + .48392;
J.LC+R: (.004401*L) + (.002266*R) .20827;
Comb.LC+R: (.004212*L) + (.001442*R) + .10362;
FMS.T: (.002279*T) + .53083; J.T: (.003376*T) +.20827;
Comb.T: (.002881*T) + .09802.

The "FMS" equations reflect regression results in the total non-Japanese sample (N = 108). "Comb" equations are based on data for the combined non-Japanese and Japanese samples

(N = 393); "J" equations are those based on data for Japanese examinees only (N = 285), reported earlier.

28. In these exploratory analyses, the primary purpose is to assess the general degree of agreement between observed LPI levels in these samples and levels estimated from alternative TOEIC/LPI linkage equations. Accordingly, emphasis is on the absolute magnitudes of the mean residual values rather than on the direction of divergence of a particular sample's average LPI performance from expectation based on TOEIC scores.

29. Note that in the Saudi sample, the mean residual associated with the "Set A" equation developed in the FMS calibration sample was comparatively small (/.15/). Moreover, mean residuals associated with FMS-calibrated equations tended to be smaller than those associated with other calibration-sample equations. These results appear to be attibutable primarily to the fact that the three non-Japanese samples were better "fitted" by equations influenced solely by FMS data than by equations reflecting the influence of the larger Japanese sample. Comparisons based on combined-sample calibration equations appear to be more pertinent in the present context, than comparisons based on FMS equations.

30. Saudi nationals earn higher means on TOEFL Listening Comprehension (LC) than on TOEFL Reading Comprehension and Vocabulary (RC&V)). For example, data for more than 20,000 Saudi nationals indicate a TOEFL-LC mean of 48 (38th percentile in a basic TOEFL reference group) and a TOEFL-RC&V mean of 42 (14th percentile). This pattern holds for Arabic speakers generally in the TOEFL testing context (e.g., ETS, 1983: 25).

Section IV

31. The educated, adult ESL users/learners likely to be tested with the TOEIC in places of work, or work-related ESL training, plausibly are the "business-oriented" counterparts of their comparably educated, but "academically-oriented" fellow nationals who take the TOEFL in conjunction with plans to study in the U.S or Canada. To the extent that this is true, trends across national samples of TOEIC examinees with respect to patterns of TOEIC performance are likely to parallel those that have been observed for corresponding national samples of TOEFL examinees. We have seen, for example, that the sample of Saudi TOEIC examinees in this study had considerably higher means on TOEIC-LC than on TOEIC-R, and that a pattern of higher means on TOEFL-LC than on TOEFL Reading Comprehension & Vocabulary is characteristic of Saudi (and other Arabic-speaking) TOEFL examinees (ETS 1985a). As data from TOEIC-use settings in various accumulate, it will be possible to assess the extent to which trends observed for national samples of TOEIC examinees parallel those that have been observed in corresponding national samples of TOEFL examinees over the past decade or more
.
32. The sample studied by Bachman and Palmer (1983) was made up of 75 Taiwanese ESL users/learners in an academic setting in the U.S. Sample means were not reported for the variables, including the interview ratings. Bachman and Palmer (1983) concluded, in part, as follows: ". . . (W)e feel we have found evidence demonstrating both the convergent and the discriminant validity of the FSI oral interview. . .; (and) have demonstrated strong support for the

distinctness of speaking and reading as traits" (p. 168).

33. Results for the Japanese testing context may be thought of as representing a relatively fully developed set of "subpopulation specific" guidelines for inferring LPI performance not only from TOEIC-LC score, but also from TOEIC-Total, and from TOEIC-R only (cf. data in "expectancy tables" based on data for 285 Japanese examinees--Tables 6.1, 6.2, and 6.3, above).

34. Strictly speaking, it is necessary at this juncture to add a qualifying assumption, namely, "assuming levels of 'reproducibility' of LPI-criterion ratings (inter-rater agreement in rank order and level) comparable to the levels sustained by the TOEIC-related interviewers/raters in this study." This is a critical assumption because, as noted by Lowe (1987: 46), the oral language proficiency interview " . . . is not an instrument because the procedure is neither fixed in print nor invariable. The procedure varies with the ability of the examinee and the skill of the interviewer(s), which represents both a strength and a weakness." Although inter-rater reliability was not directly at issue in this study, the consistently high levels of TOEIC/LPI correlations and the results of the residual analyses provide strong indirect evidence that the TOEIC-related interviewers/raters who generated the criterion data used in this study were able to sustain a high degree of reproducibility in the assignment of examinees to FSI levels. Results of an incidental analysis of inter-rater reliability in one setting, involving data for the sample of 42 Mexican examinees, are described in Appendix B: for two raters, r = .90; mean LPI levels were 1.71 and 1.66; very close agreement is indicated.

35. The distribution of LPI ratings in French or Spanish shown in the figure for the U.S.students, reflects the level of oral language proficiency attained by college seniors majoring in these languages during the 1960's. It may not be descriptive of levels of proficiency for the current generation of college majors in these languages in the U.S. Due to increased curricular emphasis on the development of productive skills since (and in part as a consequence of) the Carroll (1967) study, today's college-senior-level foreign language majors may exhibit somewhat higher levels of proficiency than their predecessors. At the same time, judging from the distribution of LPI ratings for public-secondary-school teachers of these languages in 1985--recruited largely from more recent cohorts of college language majors--relatively few of today's college-senior-level language majors are likely to be functioning above LPI Level 3--unless they happen to be native-speakers of a target language. For present purposes, it is useful to recall (from Section II, Figure 3, and related discussion) that many, if not most of the teachers rated above LPI Level 3 probably were native speakers of the languages involved--that is, they were not native-English speakers who specialized in the study of these languages from secondary-school through graduate-school in preparation for careers as language teachers or other language specialists.

36. Background questions of this type are routinely included in test admninistrations conducted under the auspices of the TOEIC Steering Committee in Japan. As noted by Saegusa (1989), there are systematic differences in TOEIC-score levels for examinees classified according to educational level, type of position, extent of use of English at work, time spent in an English-speaking environment, and so on--see Appendix C for illustrative findings from Saegusa (1989), based on data supplied by the TOEIC Steering Committee in Japan.

37. Self-ratings of oral English proficiency could be used as a surrogate for actual LPI ratings in research concerned with identifying nontest correlates of LPI performance. For example, the data from Saegusa (1989) shown in Appendix C, Exhibit C.1, indicate that examinees who reported having lived in an English-speaking country for six months or more had markedly higher TOEIC scores than did their counterparts without such experience. By regressing self-assessments of oral English proficiency on TOEIC scores and this experience variable (nominally coded) it would be possible to obtain information regarding the extent to which the experience variable contains unique criterion-related variance--that is, variance that is not already reflected in the TOEIC scores of examinees. It is plausible that the patterns of association observed in analyses involving self reports will tend to parallel those observed in analyses involving actual criterion data. This can be evaluated in research involving both self reports of oral English proficiency and actual LPI ratings as criteria.

38. According to Saegusa (1983: 101-102), "(In English-language instruction in Japan), probably listening has been the most neglected of the four language skills. . . It would be no exaggeration to say that almost all the time and energy of English education in Japan has been spent for reading comprehension." In the circumstances, it seems likely that Japanese TOEIC examinees may be more advanced in reading than in listening comprehension (and, by inference, conversational skill).

39. Problems involved in the direct assessment of writing ability are generic--that is, they are not unique to the assessment of writing ability in ESL users/learners. See Breland (1983), for a comprehensive, detailed analysis of the numerous problems that are involved in the direct assessment of writing ability in samples of U.S. college-bound high-school seniors; see also Breland (1977). To obtain a useful overview of many of these problems as they apply to the development of writing samples suitable for use in testing ESL users/learners in the TOEFL testing context, see Carlson, Bridgeman, Camp, and Waanders, 1985). The rating (scoring) procedures adopted by these investigators were not designed to classify writing samples according to the extent to which the written products involved resembled those of educated native-English-speaking (graduate-student) counterparts in the United States. At the same time, there were very sharp average differences in distributions of rated writing ability: means were 20.5 and 12.6 for native-speakers (U.S) and nonnative-speakers, respectively, on the arbitrarily defined scale employed.

40. Experience in the TOEFL testing context indicates that ESL users/learners with relatively low TOEFL scores have been able to perform satisfactorily in U.S. colleges and universities. For example, results of a study conducted by the American Association of Collegiate Registrars and Admissions Officers or AACRAO (1971), involving over 1,000 foreign students whose work was supported by the Agency for International Development, indicated that the students, on average, performed academically at a level comparable to that of their U.S. student-counterparts. The mean TOEFL Total score for this sample was 483. Some 71% of the institutions responding to a survey by the TOEFL testing program (ETS, 1981: 19) indicated that individuals with TOEFL scores below 499 would be referred to "a full-time English language program." Similarly, Campbell (1986: 61) observed that according to ESL placement test results many of the more than 1,000 foreign students tested annually at his institution (the University of

California at Los Angeles) " . . . should be spending about two-thirds of their time studying English. But instead of this they take other courses (in which, they perform well, judging from their grades). So people with major English language deficiencies may do very well in their academic programs. They do this by using various compensatory strategies (involving greater time on task)." Greater time on task may not be a viable compensatory strategy in the business world. Still, the experience cited here illustratively points up the difficulties involved in efforts to identify minimum working proficiency levels in second-language assessment contexts.

Section V

41. Staff members for the Carroll (1967) study were John L. D. Clark, Thomas M. Edwards, and Fannie A. Handrick.