ETS TOEFL®

Monograph Series

MS - 27
February 2005

Effects of Language of Administration on a Self-Assessment of Language Skills

Carsten Roever

Donald E. Powers

Effects of Language of Administration on a Self-Assessment of Language Skills

Carsten Roever
University of Melbourne, Victoria, Australia

Donald E. Powers
ETS, Princeton, NJ

RM-04-06

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org

Web site: www.ets.org/toefl

Foreword

The TOEFL Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language™ (TOEFL®) test development efforts. As part of the foundation for the development of the next generation TOEFL test, papers and research reports were commissioned from experts within the fields of measurement, language teaching, and testing through the TOEFL 2000 project. The resulting critical reviews, expert opinions, and research results have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project is a broad effort under which language testing at Educational Testing Service® (ETS®) will evolve into the 21st century. As a first step, the TOEFL program recently revised the Test of Spoken English™ (TSE®) and introduced a computer-based version of the TOEFL test. The revised TSE test, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The computer-based TOEFL test, introduced in 1998, takes advantage of new forms of assessment and improved services made possible by computer-based testing, while also moving the program toward its longer-range goals, which include:

- the development of a conceptual framework that takes into account models of communicative competence
- a research agenda that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

Monographs 16 through 20 were the working papers that laid out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, described the process by which the project was to move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extended the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). These conceptual frameworks guided the research and prototyping studies described in subsequent monographs that resulted in the final test model.

As TOEFL 2000 projects are completed, monographs and research reports will continue to be released and public review of project work invited.

TOEFL Program Office
Educational Testing Service

Abstract

Self-assessments of English language skills have proven useful in a variety of settings. One threat to the validity of such assessments, however, is that responses may differ in meaning according to whether the assessment is administered in English or in the respondent's native language. This study investigated the effect of administering a self-assessment of English language skills in English versus in the self-assessor's native language. Study participants—115 volunteers located at test sites in Germany, Mexico, Korea, and Taiwan—completed self-assessments in both their native languages and English. The results revealed comparable responses in both languages in terms of reliability, level, and variation. Most important, the correlations between self-assessment scales given in English and those given in participants' native languages were virtually perfect when corrected for attenuation due to unreliability.

Key words: Self-assessment, score anchoring, new TOEFL®, validation, questionnaire

Table of Contents

List of Tables

Page

Introduction

This study entailed the validation of a self-assessment questionnaire of academically oriented English as a second language (ESL) proficiency for a TOEFL-like population. The research was undertaken as part of the development of score descriptors based on ability statements ("can-do" statements) for a revised Test of English as a Foreign Language™ (TOEFL®). Because one source for these descriptors was test-taker responses to a self-assessment instrument, our concern was whether the administration of self-assessment scales in English, rather than in test takers' native languages, would interfere with the accuracy of the responses. This interference would attenuate the validity of score descriptors and thus the inferences that test users might draw from them.

We first briefly review some evidence of the usefulness of self-assessment in language testing. Next we discuss selected issues related to the language of administration.

Self-Assessment in Language Testing

Self-assessments have proven useful in a variety of contexts. There is substantial support, both conceptually and empirically, for the belief that self-assessors frequently have both the information and the motivation to make effective judgments about their own behaviors (Shrauger & Osberg, 1981). This seems to be especially true for language learners (Upshur, 1975).

Many of the issues relevant to self-assessment in general are germane to the self-assessment of language skills. There is at least one notable difference, however. As LeBlanc and Painchaud (1985) noted, adult second-language learners have already mastered at least one language and thus know what it means to be able to use a language. They are, therefore, often especially well-positioned to evaluate how they are performing. However, because self-assessors may be prone to overestimation and because self-assessments may lack comparability across cultures, some researchers (e.g., Davidson & Henning, 1985) have found little reason to trust student ratings of second-language skills, concluding that the "phenomenon of exaggeration in ability estimation may be an inherent weakness of self-reports of language abilities" (p. 175). Other researchers, however, have been decidedly more positive about the potential of language self-assessments. This optimism is consistent with a number of empirical studies.

The results of validity studies of language self-assessments are varied but, on balance, positive. Heilenman (1990) computed a correlation of .33 between course grades and undergraduate students' self-assessments of their French language skills (grammar, vocabulary,

accuracy, and fluency). Clark (1981) compared self-assessments of speaking, reading, listening, and writing with FSI (Foreign Service Interview) scores and reading and listening test scores, finding correlations of almost .60. LeBlanc and Painchaud (1985) found correlations of .80 and .82 between proficiency test scores and self-assessments of reading, listening, writing, and speaking skills. Working some years later in the same context, however, Bayliss (1991) was unable to reproduce the results obtained by LeBlanc and Painchaud.

Wilson (1999) administered the Test of English for International Communication™ (TOEIC®) to some 900 Swiss workers enrolled in ESL training programs and collected self-ratings that were translated and administered in the participants' native languages (either French or German). Wilson found that self-ratings correlated .75 and .70 with TOEIC listening and reading scores, respectively. Tannenbaum, Rosenfeld, Breyer, and Wilson (2000) administered the TOEIC and 75 can-do statements (15 each in L, R, S, W, and interactive (S/L) to some 8,000 examinees). These investigators found correlations of TOEIC reading with self-ratings across the five domains to be .65.

Ross (1998) conducted a meta-analysis of studies dealing with self-assessment in second and foreign languages. For reading, he located 23 correlations, with an average $r = .61$. In another summary of research on second-language self-assessments, Oscarson (1997) concluded, among other things, that

- "a clear majority" of studies report positive results
- accuracy depends to a considerable degree on the purpose of the assessment
- assessments are more accurate when based on task content that is closely related to students' situations as potential users
- assessments are more accurate when they are administered in the subjects' native languages

It is Oscarson's last observation that motivated the study reported here, which was designed to evaluate the influence of one potential source of construct-irrelevant variance in self-assessments of English as a second language. The source of concern here was the language in which the self-assessment is administered. Our hypothesis was that respondents might interpret or respond differently to self-assessments when they are administered in English rather than in the self-assessors' native languages.

Language Effects in Assessment Instruments

With the increasing use of internationally used assessments has come greater interest in the effects of the language in which an assessment is administered. The discussion has centered mainly on test translation, for which guidelines have now been developed and finalized by the International Test Commission (ITC) (Hambleton, 2001). These guidelines discuss test context (comparability of constructs between populations), test development (appropriateness of the test format), test administration (operational factors), and documentation/score interpretation (comparisons between populations).

Though informative, the ITC guidelines are only partially relevant to our main concern, however—that is, the extent to which administering questionnaires in English to speakers of other languages may compromise the validity of inferences based on them. The two major factors that might attenuate the validity of an English language self-assessment instrument are (a) language comprehension problems and (b) nonequivalence of constructs between groups.

To some extent, language comprehension problems are almost unavoidable when administering a questionnaire to nonnative speakers. However, we speculated that a lack of comprehension would be less problematical here because (a) the typical TOEFL test taker possesses an intermediate to advanced level of English proficiency and (b) the instrument of interest was written in "simple English."

Nonequivalence of constructs is a more serious and perhaps intractable problem. For example, applying the ITC guidelines to a math test adapted for Chinese students, researchers found that the content domain from which the test was sampled was not equivalent to the content covered in the participating Chinese schools (Hambleton, Yu, & Slater, 1999), and comparisons between populations had to be restricted.

In another study concerned with construct equivalence, Whitworth (1988) administered a personality test battery (the Minnesota Multiphasic Personality Inventory or MMPI) to monolingual English speakers (who took the battery in English), monolingual Spanish speakers (who took the battery in Spanish), and bilingual English-Spanish speakers (who took the battery in their preferred language). Whitworth found that subjects taking the test in Spanish scored higher on 9 of 13 scales than did bilingual English-Spanish speakers taking the test in English and higher on 10 of 13 scales than did monolingual English speakers. Bilingual speakers taking the test in English scored higher than English monolinguals on only one scale. Whitworth

speculated that, being more acculturated, the bilinguals who took the English version adhered more closely to cultural and behavioral norms of the English-speaking population than did the participants who took the Spanish version.

An alternative explanation is that different test languages trigger different behavioral patterns in bilinguals: When tested in Spanish, their behaviors conform to the behavioral norms of the Spanish-speaking reference groups, whereas when tested in English, they conform to the norms of the English-speaking group. Because none of Whitworth's subjects took the battery in both languages, it is impossible to discount either of these explanations. However, his findings show the very real possibility that the language in which a questionnaire is administered can strongly affect responses and, as a result, may cast doubt on any inferences drawn from them.

To reiterate, the main objective of this study was to investigate the degree to which self-assessors may respond differently to self-assessments when they are administered in English rather than in the self-assessors' native languages.

Method

Materials/Procedure

Study participants were asked to complete a biographical background information form, an experimental 60-item validity questionnaire, and the self-assessment questionnaire in English and in the participants' native languages. Before giving their consent to participate in the study, students were told that the purpose of the study was to examine the relationship between self-assessments and abilities in English for academic purposes.

The self-assessment questionnaires consisted of nine scales (Table 1), seven of which were employed in this study. The first four scales (five can-do statements for each) related to listening, writing, speaking, and reading skills. An example of a statement for listening is the following:

I can understand the main ideas of lectures and conversations.

Participants assessed themselves on a 5-point scale ranging from "Extremely well" to "Very poorly."

Two other scales required participants to rate their English language ability for each of the four skills in comparison with that of fellow students in ESL classes (language class comparison) and content classes taught in English (content class comparison). The 5-point scale

ranged from “A lot higher” to “A lot lower.” A final scale required rating of overall language ability for each of the four skills on a 5-point scale from “Extremely good” to “Poor” (skill rating). Only items that employed the same response format were combined to form composite scales. The stems for all questions are shown in the appendix.

Table 1

Scales, Example Statements, and Rating Categories

Scale	Name	Example statement (scale 1-4) / Instructions (scales 5-7)	Rating descriptors
1	Listening	I can understand the main ideas of lectures and conversations	Extremely well – Very well – Well – Somewhat poorly – Very poorly
2	Writing	I can write an essay in class on an assigned topic	"
3	Speaking	I can speak for one minute in response to a question	"
4	Reading	I can understand vocabulary and grammar when I read	"
5	Language class comparison	How does your English language ability compare to the ability of other students in your classes?	A lot higher – Somewhat higher – About the same – Somewhat lower – A lot lower
6	Content class comparison	How does your English language ability compare to the ability of other students in your classes?	"
7	Skill rating	Please rate your overall English language ability in each of the four skills.	Extremely good – Very good – Good – Not very good – Poor

Translators and Translations

For this study, we selected as target languages those having large representations in the TOEFL population—Chinese, Spanish, German, and Korean. For each of the first three target languages, we identified two individuals who could serve as translators. Questionnaires were translated independently by both translators, who then conferred to produce a final, unified version. For the fourth language (Korean), we identified only one translator. Because the Spanish language questionnaire was to be administered in Mexico and the Chinese language questionnaire in Taiwan, familiarity with South American Spanish and classical Chinese characters was required. The translator teams consisted of the following individuals:

German—a native speaker of high German and a native speaker of Austrian German, both with high proficiency in American English

Spanish—a native speaker of Chilean Spanish and a native speaker of German with near-native proficiency in South American Spanish and extensive translator experience for English-South American Spanish translation (this translator was also the Austrian German translator in the German translation team)

Chinese—two native speakers of Chinese from Taiwan with high proficiency in American English

Korean—one native speaker of Korean with high proficiency in American English

For reasons of time and expense, we did not incorporate back-translation.

Data Collection Procedures

Only the instruments described below were administered to participants. Data were collected independent of the TOEFL administration.

The main study. Study coordinators were identified at the test sites in Germany, Mexico, Korea, and Taiwan. These coordinators recruited study participants whose first language was the local language and who were enrolled at a local university or (in the case of Taiwan only) a language school. Participants received the local equivalent of U.S.$25 upon completing the assessment. Participants first completed a background questionnaire and then, in a group session, the self-assessment questionnaires in the following, fixed sequence:

1. self-assessment questionnaire *in English*
2. an experimental 60-item validity questionnaire (not discussed here)
3. self-assessment questionnaire *in the participants' native languages*

The 60-item validity questionnaire was administered for a separate, related research study. It also served our purposes here by temporally separating the English and native language administrations of the self-assessment questionnaire so respondents would be less likely to simply remember and repeat the same responses for both language versions. We had no reason to believe that the administration of the intervening validity questionnaire would affect responses to the subsequent self-assessment. The order of administration (i.e., the English language version before the native language version) was chosen so as to preclude the possibility that, because participants would be likely to understand the questions, they might also *remember* their responses and simply repeat them when given the presumably more difficult-to-understand English language version of the questionnaire. To us, this factor outweighed the potential benefits of employing a counterbalanced administration to detect any effects of the order of questionnaire administration. Such a design might have, we believed, masked any language-of-administration effects.

The comparison study. Study coordinators were identified at two test sites in the United States. These coordinators recruited study participants whose first language was a language other than English and who were enrolled at a local university. Participants had either recently taken TOEFL or were considered by coordinators to be "ready" to take TOEFL. Participants were paid U.S.$35 (instead of $U.S.25) upon completing the assessment, as the intervening questionnaires in this study required more time than those used in the main sample. Participants first completed a background questionnaire and then, in a group session, the self-assessment questionnaires in the following, fixed sequence:

1. self-assessment questionnaire *in English*
2. an experimental 60-item validity questionnaire (not discussed here)
3. a second experimental 60-item validity questionnaire (not discussed here)
4. self-assessment questionnaire *in English* (identical to the first questionnaire)

The 60-item validity questionnaires were administered for a separate, related research study. We readministered the self-assessment questionnaire in English simply to assess the degree to which participants could be expected to give consistent responses to a questionnaire of this sort. The results of this test-retest administration would therefore serve as a baseline reading against which we could judge the level of agreement between English and native language versions of the self-assessment.

Respondents

Main study. In total, 115 participants responded to our questionnaires, all of them university students (Korean, Mexican, and German samples) or language school students (Taiwanese sample). The characteristics of the responding sample were as follows. Most (68%) were female. Participants ranged in age from 14 to 53 years, with a mean and median of 21 years. Their native languages were Chinese (27%), Korean (26%), Spanish (25%), German (16%), and several other languages (Czech, Russian, Romanian, English, Hungarian), which were excluded from the analysis, leaving a sample of 106 participants. A slight majority (54%) reported having taken TOEFL before, and 30% reported previous TOEFL scores (mean of 554 on the paper-and-pencil TOEFL scale, which converts to 217 on the computer-based TOEFL scale).

Comparison study. A smaller sample of 72 participants took the self-assessment instrument in English only in a test-retest design with two versions of the experimental 60-item validity questionnaire intervening. All participants were university students in the United States or Canada, and most of these participants were female (57%); their ages ranged from 19 to 39 years, with a median age of 26 years. The native languages of these participants were Chinese (39%), Korean (20%), and Farsi (12%). Sixteen other languages were each represented by fewer than three speakers. All of these participants had been recruited at universities in the United States and Canada, and the length of their stay ranged from 1 month to 6 years, with a median of 1 year.

Data Analysis

The questionnaire responses were analyzed with SPSS version 10.1. Descriptive statistics were computed for each item and each scale. Scales were correlated, and an ANOVA was

computed for the main sample to identify scales for which test-taker self-assessments differed significantly between L1 groups.

Results

Table 2 shows the average rating on each of the self-assessment subscales when administered in English and in participants' native languages (L1).

Table 2

Means and SDs on Each Self-Assessment Scale by Language of Assessment Administration

	N	Mean	*SD*
Listening English	105	3.35	.69
Listening L1	104	3.40	.69
Writing English	106	3.05	.65
Writing L1	105	3.06	.72
Speaking English	106	3.32	.73
Speaking L1	106	3.29	.74
Reading English	106	3.52	.71
Reading L1	106	3.43	.71
Lang. class comparison English	99	3.32	.60
Lang. class comparison L1	99	3.31	.60
Content class comparison English	78	3.37	.65
Content class comparison L1	78	3.32	.68
Overall English	105	3.05	.60
Overall L1	106	3.21	.61

Each of the differences between responses given in English versus in the native language is small, and none is statistically significant. Moreover, the standard deviations for each version are comparable for each of the responses.

Table 3 shows the average self-ratings on the first and second ratings for participants who retook the questionnaire in English in the comparison study.

Table 3

Means and SDs on Each Self-Assessment Scale for First and Second Administrations

	N	Mean	*SD*
Listening	68	3.81	.65
Listening retest	68	3.78	.63
Writing	67	3.41	.61
Writing retest	67	3.42	.64
Speaking	69	3.61	.78
Speaking retest	69	3.60	.72
Reading English	66	4.07	.72
Reading retest	66	3.86	.69
Lang. class comparison	32	3.59	.58
Lang. class comparison retest	32	3.66	.56
Content class comparison	61	2.95	.79
Content class comparison retest	61	2.93	.76
Overall	68	3.43	.53
Overall retest	68	3.39	.59

Again, each of the differences between means is small and none is statistically significant. Standard deviations are also comparable.

Table 4 shows internal consistency reliabilities (coefficient alpha) for the seven subscales when administered in English and when administered in the L1, for both the interlanguage and test-retest samples.

Table 4

Internal Consistency Reliability Estimates (Coefficient Alpha)

	Interlanguage sample α	Test-retest sample α
Listening English	.88	.89
Listening L1	.89	.89
Writing English	.83	.82
Writing L1	.86	.85
Speaking English	.88	.92
Speaking L1	.90	.92
Reading English	.91	.94
Reading L1	.91	.93
Lang. class comparison English	.68	.68
Lang. class comparison L1	.63	.61
Content class comparison English	.76	.84
Content class comparison L1	.74	.82
Overall English	.63	.63
Overall L1	.60	.54

Table 5 shows the correlations between scales for the English and L1 versions (interlanguage correlation) and the test-retest correlation for the group that reanswered the questionnaire. The same correlations, corrected for attenuation, are also shown.

Table 5

Correlations Between English and L1 Versions (Interlanguage) and First and Second Administrations (Test-Retest) for Self-Assessments

	N	Inter-language correlation	Corrected for attenuation	*N*	*Test-retest correlation*	*Corrected for attenuation*
Listening	103	.81	.92	*68*	*.70*	*.79*
Writing	105	.82	.97	*67*	*.78*	*.93*
Speaking	106	.83	.93	*69*	*.84*	*.91*
Reading	106	.76	.84	*66*	*.77*	*.82*
Lang. class comparison	99	.90	>1.0	*32*	*.80*	*>1.0*
Content class comparison	76	.87	>1.0	*61*	*.91*	*>1.0*
Overall English	105	.76	>1.0	*68*	*.72*	*>1.0*

Note. All correlations are significant at the .001 level or beyond.

For each scale, the correlation between English and L1 versions are high—in most cases virtually perfect—and at least as strong as the corresponding test-retest correlation. (In some cases, the corrected correlations are greater than 1. Estimates that exceed unity are not uncommon, as the assumptions that underlie the formula for correcting for attenuation are not always fully met. Moreover, the corrections are indeed *estimates*, which can by chance sometimes exceed 1.) The conclusion from these results, therefore, is that the two questionnaire versions are reflecting the same construct.

To identify possible differences between L1 groups in the main sample (Chinese, German, Korean, and Spanish), an ANOVA was undertaken to compare responses to the English-language questionnaires and L1 questionnaires. A significant difference was detected for responses to L1 questionnaires for listening self-assessment ($F = 3.07$, $df = 103$; $p = .031$), and for English-language questionnaires, for comparison with fellow students in language classes ($F = 3.21$, $df = 98$; $p = .026$), and for self-assessment of overall L2 proficiency ($F = 3.86$,

df = 104; $p = .012$). Scheffé posthoc tests confirmed that (a) the German group assessed itself significantly higher than the Chinese group on the overall self-assessment rating in English ($p = .042$) and (b) the Mexican group assessed itself significantly higher in comparison with fellow students in language classes than the German group ($p = .038$). No significant differences between groups were found for listening self-assessment in the L1.

Discussion

The results suggest that administering a self-assessment questionnaire in English rather than in the respondent's L1 may have little if any effect on responses for a population similar to typical TOEFL test takers. It seems unlikely therefore that presenting a self-assessment questionnaire in English will attenuate the validity of inferences about respondents' English language skills.

Our conclusion is based on several findings. First of all, with few exceptions, the reliability coefficients of the English questionnaire subscales were generally quite good, regardless of the language (English or L1) in which they were administered. More important, the various scales exhibited comparable reliability for the English and native language versions of the instruments.

Second, the interlanguage correlations were very similar to, and in some cases higher than, the retest correlations, which served as an appropriate baseline against which to gauge them. The small, nonsystematic differences between correlations leads us to believe that the interlanguage correlations are essentially comparable to the retest correlations, which were based only on the English language version of the questionnaires and were therefore free from attenuation due to differences in the language in which the questionnaire was administered. Although it is conceivable that the very small differences obtained between the English and L1 versions were due to differential levels of comprehension, the size of the differences seems to be of little practical significance.

Third, the level and variation of self-assessment scores were essentially the same for the English and L1 questionnaires and, in the retest condition, for the first and second administrations of the self-assessment questionnaire. Participants did not rate themselves systematically higher (or lower) according to the language in which the questionnaire was administered (in English or in the native language), just as they did not systematically rate themselves higher on the first or second administration of the original questionnaire.

Finally, and most importantly, the correlations between self-assessment ratings in English and in the L1 were quite high, ranging from .76 to .90 on the seven subscales. When corrected for attenuation, the correlations were near perfect. This indicates that participants activated their knowledge about their own language competence similarly when responding in English and in their L1.

A limitation of this study is its relatively narrow target population. We sampled only from a population of students who had previously taken TOEFL or who were judged by their teachers to be "ready" to take the test. Because TOEFL is targeted at medium-to high-proficiency students, it is likely that differences due to questionnaire language could be larger and more problematic for lower proficiency groups. Therefore, our findings should not be interpreted as showing that questionnaire language does not make a difference—rather, it did not have an effect in a TOEFL-like population.

References

Bayliss, D. (1991, March). *The role and limitations of self-assessment in testing and research.* Paper presented at the Language Testing Research Colloquium, Princeton, NJ.

Clark, J.L.D. (1981). Language. In T. S. Barrows et al. (Eds.), *College students' knowledge and beliefs: A survey of global understanding* (pp. 25-35). New Rochelle, NY: Change Magazine Press.

Davidson, F., & Henning, G. (1985) A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing, 2,* 164-179.

Hambleton, R.K. (2001). *The next generation of the ITC test translation and adaptation guidelines.* Unpublished manuscript.

Hambleton, R.K., Yu, J., & Slater, S.C. (1999). Fieldtest for the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment, 15*(3), 270-276.

Heilenman, L.K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7,* 174-201.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19,* 673-687.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 175-187). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15,* 1-20.

Shrauger, J.S., & Osberg, T.M. (1981). The relative accuracy of self-predictions and judgments by others of psychological assessment. *Psychological Bulletin, 90,* 322-351.

Tannenbaum, R.J., Rosenfeld, M., Breyer, F.J., & Wilson K. (2000). *Linking TOEIC scores to self-assessments of English-language abilities: A study of score interpretation.* Unpublished manuscript.

Upshur, J. (1975). Objective evaluation of oral proficiency in the ESOL classroom. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967-1974* (pp. 53-65). Washington, DC: TESOL.

Whitworth, R.H. (1988). Anglo- and Mexican-American performance on the MMPI administered in Spanish or English. *Journal of Clinical Psychology, 44*(6), 891-897.

Wilson, K.M. (1999). *Validity of a global self-rating of ESL speaking proficiency based on an FSI/ILR-referenced scale* (ETS RR-99-13). Princeton, NJ: ETS.

Appendix

Self -Assessment Questions

I can understand the main ideas of lectures and conversations.

I can understand important facts and details of lectures and conversations.

I can understand the relationships among the ideas in a lecture.

I can understand a speaker's attitude or opinion about what he or she is saying.

I can recognize why a speaker is saying something (for example, to explain something, to complain about something, to agree with someone.)

I can write an essay in class on an assigned topic.

I can summarize and paraphrase in writing information that I have **read** in English.

I can summarize in writing information that I have **listened to** in English.

I can use correct grammar, vocabulary, spelling, and punctuation when I write in English.

I can speak for one minute in response to a question.

When I speak in English, other people can understand me.

I can participate in conversations or discussions in English.

I can orally summarize information from a talk I have **listened to** in English.

I can orally summarize information I have **read** in English.

I can understand vocabulary and grammar when I read.

I can understand major ideas when I read.

I can understand how the ideas in a text relate to each other.

I can understand the relative importance of ideas when I read.

I can organize or outline the important ideas and concepts in texts.

Note. Responses were "Extremely well," "Very well," "Well," "Not very well," and "Not at all."

If you are taking classes in English, please complete this table. How does your English language ability compare to the ability of other students in your classes:

Skill:

Reading
Listening
Speaking
Writing

Note. Responses were "A lot higher," "Somewhat higher," "About the same," "Somewhat lower," and "A lot lower."

If you are studying a subject in English (for example, biology or business), please complete this table. How does your English language ability compare to the ability of the other students in your classes?

Skill:

Reading
Listening
Speaking
Writing

Note. Responses were "A lot higher," "Somewhat higher," "About the same," "Somewhat lower," and "A lot lower."

Please rate your overall English language ability in each of the four skills.

Skill:

Reading
Listening
Speaking
Writing

Note. Responses were "Extremely good," "Very good," "Good," "Not very good," and "Poor."

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

Email: toefl@ets.org

Web site: www.ets.org/toefl

* America Samoa, Guam, Puerto Rico, and US Virgin Islands

I.N. 990116