TOEFL The Researcher

Test of English as a Foreign Language • Princeton, NJ • USA Fall 2005

Discourse Characteristics of Test Takers' Spoken and Written Responses

Among the most innovative features of the TOEFL® Internet-based test (TOEFL iBT) are the addition of a speaking measure and a revised writing measure consisting of two writing tasks. The speaking and writing measures include both independent and integrated tasks. Independent tasks require test takers to talk or write about their personal opinions or experiences. Integrated tasks require the use of more than one language skill. For example, one writing task requires test takers to read a short passage, listen to a brief lecture, and then write about how the information in the passage and lecture are related.

The rationale for including tasks that require speaking and writing on TOEFL iBT is that these tasks directly assess test-takers' abilities to use English to communicate in academic situations. An assumption underlying this rationale is that these abilities should be evident in the characteristics of the discourse the test takers produce when completing these tasks. Analyses of discourse characteristics provide an important source of validity evidence for the interpretation of TOEFL iBT scores.

As new task types were being developed for TOEFL iBT, two groups of researchers investigated how the discourse characteristics of spoken and written responses varied with ability level and task type.

Using responses to early prototypes of speaking tasks, Brown, Iwashita, McNamara, and O'Hagan (2005) analyzed the nature of the discourse characterizing speech samples at different levels of ability and for different task types. These researchers first identified the categories of speech qualities that English as Second Language (ESL) experts attended to when rating the responses. Brown et al. then analyzed aspects of test-taker discourse in the responses associated with the experts' categories, such as linguistic resources, phonology, fluency, and content. Many of the features—grammatical accuracy and complexity, vocabulary, pronunciation, speech rate, amount of content—in each of the categories were found to vary with score level to some degree, and a lesser amount with task type. When compared with performances on independent tasks, performances on integrated tasks had a more complex schematic structure and included more sophisticated vocabulary, but were less fluent. Their finding of more complex schematic structure for responses to integrated tasks was consistent with the rationale for including such tasks.

In a similar manner, Cumming, Kantor, Baba, Eouanzoui, Erdosy, and James (2005) reported important differences in the discourse characteristics of written responses related to proficiency level as well as task types. Greater writing proficiency was associated with longer responses, greater lexical sophistication, syntactic complexity, and grammatical accuracy. Cumming et al. also compared the qualities of responses to the two task types. Responses to independent tasks were longer, had a more fully developed argument structure, relied on the self as a source of evidence, and used declarations—statements of personal opinions or facts. Responses to the integrated tasks had greater lexical sophistication and syntactic complexity, relied on the source materials for information, and used paraphrasing and summarization.

To sum up, these researchers documented how the discourse characteristics of spoken and written responses to prototype test tasks varied with proficiency level and with task type. These studies have two important implications. First, holistic scores on speaking and writing tasks can be verified and anchored empirically through the analysis of important discourse qualities that differentiate proficiency levels and task types.

Second, the introduction of integrated tasks on TOEFL iBT would have the desired effect of broadening the constructs assessed by the speaking and writing measures. The integrated tasks provided test takers an opportunity to demonstrate discourse qualities that were different from those demonstrated on independent tasks.

References

Brown, A., Iwashita, N., McNamara, T., & O'Hagan, S. (2005). *An examination of rater orientations and test-taker performance on English-for-academic purposes speaking tasks.* (TOEFL Monograph Series No. MS-29, ETS Research Report RR-05-05). Princeton, NJ: ETS.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (in press). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for next generation TOEFL* (TOEFL Monograph Series). Princeton, NJ: ETS.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10,* 5-43.

TOEFL Reports* — Areas of Inquiry

AREA	TOEFL				TSE/ SPEAK	TWE	TOEFL 2000
	GENERAL	LISTENING	STRUCTURE	READING			
TEST VALIDATION							
Construct Validity	5, 6, 10, 12, 16, 17, 21, 27, 28, 32, 36, TR1, TR5, TR11	5, 6, 10, 12, 16, 17, 20, 21, 27, 28, 32, 33, 34, 36, 51, 79, TR1, TR5, TR11	5, 6, 10, 12, 16, 17, 21, 27, 28, 32, 36, TR1, TR5	5, 6, 10, 12, 16, 17, 21, 27, 28, 32, 35, 36, 44, 47, 53, TR1, TR5, TR11	4, 7, 13, 36, 45, 48, MS7, MS9	19, 36, 38, 64	MS1, MS4, MS5, MS6, MS8, TR 14, MS10, MS21, MS25
Face/Content Validity	1, 16, 17, 21	1, 2, 16, 17, 20, 21, 71	1, 16, 17, 21, 71				
Predictive Validity	10, 16, 41	10, 16, 20, 41, 71	10, 16, 41, 71	10, 16, 41, 71	7, 13, 49, 63		
Concurrent Validity	3, 5, 10, 12, 16, 69	3, 5, 10, 12, 16, 33, 69	3, 5, 10, 12, 16, 69	3, 5, 10, 12, 16, 35, 69	4, 7, 48, 49, 58	19	MS26, MS27, MS28, MS31, MS32
TEST INFORMATION							
Score Interpretation	3,5,10,12,36,41,81,TR11	3, 5, 10, 12, 36, 41, TR11	3, 5, 10, 12, 36, 41, TR11	5, 10, 12, 36, 41, TR11	36	36, 38	MS3
Underlying Processes	36	33, 36	36	36	36, 74	36	
Diagnostic Value	27	27	27	27		67	
Performance Descriptors	41	41	41	41			
Reporting/Scaling	27, TR1, TR2	27, TR1, TR2	27, TR1, TR2	27, TR1, TR2	48, 58	38, 52, 55	TR13
EXAMINEE PERFORMANCE							
Difference Variables	1, 3	1, 3	1, 3	1, 3, 25		50, 72, 75, 76, 77	MS23
Language Acquisition/Loss	45	45	45	45			
Sample Dimensionality	28, TR5	28, TR5	28, TR5	28, TR5			
Person Fit						38, 52, 55	
TEST USE							
Decisions/Cut Scores	1, 2, 16, 57	1, 2, 16, 57	1, 2, 16, 57	1, 2, 16, 57	13		57
Test/Item Bias	9, 29, 61	9, 29	9, 29	5, 9, 11, 16, 57			61
Sociological/Pedagogical Impact	14, 59	14	14	14, 25			MS15
Satisfying Assumptions	30, TR6	30, TR6	30, TR6	30, 47, TR6, TR10			
Examinee/User Populations	5, 9, 11, 16, 57, 59, 60	5, 9, 11, 16, 57	5, 9, 11, 16, 57	5, 9, 11, 16, 57			57, 59, 60, MS14, MS15
TEST CONSTRUCTION							
Format Rationale/Selection	23, 24	23, 26, 33, 34	23, 26	23, 26, 35, 47	48	15, 54	MS18, MS19, MS20
Equating	TR2, TR4, TR7, TR8	TR2, TR4, TR7, TR8	TR2, TR4, TR7, TR8	TR2, TR4, TR7, TR8	58	52, 55	MS2
Item Pretesting/Selection	TR6, TR9, TR16, TR17	TR6, TR9	TR6, TR9	TR6, TR9, TR11, TR17		42	
Component Length/Weight				47	48	39	
TEST IMPLEMENTATION							
Testing Time	30	30	30	30, 47, TR10		39	
Scoring/Rating		TR3			4, 18, 48, 49, 65, TR5	38, 73	MS22, 70, MS28
Practice/Sequence Effects	8, 22, 61, 62	22, 24	22	22			
TEST RELIABILITY							
Internal Consistency	16, 45, TR12	16, 45, TR3, TR12	16, 45, TR12	16, 45, TR12	40	38, 42	TR12
Generalizability							MS28, MS31
Alternative Forms	45	45	45	45		42	
Test-Retest	45, TR6	45, TR6	45, TR6	45, TR6			
Inter-/Intrarater					4, 7, 18, 40, 49, TR15	19, 38, 55	
APPLIED TECHNOLOGY							
Innovative Formats	2, 23, 61, 62	2, 23, 26, 33, 34, 66, 78	2, 23, 26, 78	2, 23, 26, 35, 78			MS11, MS12, 61, 62, MS13, MS18, MS19, MS20
Machine Test Construction	TR9	TR9	TR9	TR9			
Computer-Based Testing	31, 61, 62	31, TR16	31	31, TR16			68
Item Banking	TR9	TR9	TR9	TR9			

*Research reports are identified by their series numbers; technical reports are listed by their series numbers preceded by "TR"; monographs are preceded by "MS."

For more information about TOEFL research reports and monographs, please visit: www.ets.org/toefl/research

New Research Report

RR-80. Mapping English Language Proficiency Test Scores Onto the Common European Framework

Richard J. Tannenbaum and E. Caroline Wylie

The Common European Framework describes language proficiency in reading, writing, speaking, and listening on a six-level scale. The framework provides a common language with which to discuss students' progress. This report describes a study conducted with two panels of English language experts to map scores from four tests that collectively assess reading, writing, speaking, and listening on two levels of the framework. Panel 1 recommended cut scores for the Test of English as a Foreign Language™ (TOEFL), the Test of Spoken English™ (TSE®), and the Test of Written English™ (TWE®). Panel 2 recommended cut scores for the Test of English for International Communication™ (TOEIC®). A modification of the Angoff (1971) standard-setting approach was used for multiple-choice questions, and a benchmark method (Faggen, 1994) or examinee paper selection method (Hambleton, Jaeger, Plake, & Mills, 2000) was used for constructed-response questions.

New Research Monographs

MS-28. Dependability of Scores for a New ESL Speaking Test: Evaluating Prototype Tasks

Yong-Won Lee

This study considered two critical issues concerning score dependability of the new TOEFL multitask speaking measure: How much would the score dependability be impacted by (a) combining scores on different task types into a composite score and (b) rating each task only once? To answer these questions, the study used generalizability theory (G-theory) procedures to examine relative effects of tasks and raters on examinees' speaking scores and the impact of the numbers of tasks and raters per speech sample and of subsection lengths on the dependability of speaking section scores. Univariate and multivariate G-theory analyses were conducted on rating data collected for 261 examinees for the study. The finding in the univariate analyses was that it is more efficient to increase the number of tasks rather than the number of ratings per speech sample to maximize the score dependability. The multivariate G-theory analyses revealed that universe scores among the task-type subsections were highly correlated and that slightly larger gains in composite score reliability would result from increasing the number of listening-speaking tasks for the fixed section lengths.

MS-29. An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks

Annie Brown, Noriko Iwashita, and Tim McNamara

This report documents two coordinated exploratory studies into the nature of oral English-for-academic-purposes (EAP) proficiency. Study I used verbal-report methodology to examine field experts' rating orientations, and Study II investigated the quality of test-taker discourse on two different Test of English as a Foreign Language (TOEFL) task types (independent and integrated) at different levels of proficiency. Study I showed that, with no guidance, domain experts distinguished and described examination. This study examines the impact of various qualitatively different performances using a common set of criteria very similar to those included in draft rating scales developed for the tasks at ETS. Study II provided empirical support for the criteria applied by the judges. The findings indicate that raters take a range of performance features into account within each conceptual category and that holistic ratings are driven by all of the assessment categories rather than, as has been suggested in earlier studies, predominantly by grammar.

MS-30. Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for Next Generation TOEFL

Alister Cumming, Robert Kantor, Kyoko Baba, Keanre Eouanzoui, Usman Erdosy, and Mark James

We assessed whether and how the discourse written for prototype integrated tasks (involving writing in response to print or audio source texts) field tested for the next generation TOEFL test differs from the discourse written for independent essays (i.e., the TOEFL essay). We selected 216 compositions written for six tasks by 36 examinees in a field test—representing Score Levels 3, 4, and 5 on the TOEFL essay—and coded the texts for lexical and syntactic complexity, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source text. Analyses with nonparametric MANOVAs showed that the discourse produced for the integrated writing tasks differed significantly at the lexical, syntactic, rhetorical, and pragmatic levels from the discourse produced in the independent essay. In certain analyses, these differences were also obtained across the three ESL proficiency levels.

MS-31. Dependability of New ESL Writing Test Scores: Evaluating Prototype Tasks and Alternative Rating Schemes

Yong-Won Lee and Robert Kantor

Possible integrated and independent tasks were pilot tested for the writing section of the next generation TOEFL

MS-32. Factor Structure of the LanguEdge™

rating designs as well as the impact of the number of tasks and raters on the reliability of writing scores based on integrated and independent tasks from the perspective of generalizability theory (G-theory). Both univariate and multivariate G-theory analyses were conducted. It was found that (a) in terms of maximizing the score reliability, it would be more efficient to increase the number of tasks rather than the number of ratings per essay; (b) two particular single-rating designs having different tasks for the same examinee rated by different raters [p × (R:T), R:(p × T)] achieved relatively higher score reliabilities than other single-rating designs; and (c) a somewhat larger gain in composite score reliability was achieved when the number of listening-writing tasks was larger than the number of reading-writing tasks.

Test Across Language Groups

Larry J. Stricker, Don A. Rock, and Yong-Won Lee

This study assessed the factor structure of the LanguEdge test and the invariance of its factors across language groups. Confirmatory factor analysis of individual tasks and subsets of items in the four sections of the test, Listening, Reading, Speaking, and Writing, was carried out for Arabic-, Chinese-, and Spanish-speaking test takers. Two factors were identified, Speaking and a fusion of the other sections of the test. The number of factors, the factor loadings, and the factors' error variances were invariant in the three samples, although the correlations between the factors differed. The failure to find separate factors for each section of the LanguEdge test necessarily raises questions about the test's functioning that need to be resolved.

Additional Research Reports

RR–1. **The Performance of Native Speakers of English on the Test of English as a Foreign Language.** Clark. November 1977. Discusses the results of forms of the TOEFL test administered in 1977 to native speakers of English just prior to their graduation from a college-preparatory high school program; reinforces earlier findings that the TOEFL test is not psychometrically appropriate for native speakers of English.

RR–2. **An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.** Pike. June 1979. Describes an extensive research study conducted from 1972 to 1974 that was designed to explore possible changes in the format and content of the TOEFL test; contributed to the restructuring of the test beginning in 1976.

RR–3. **The Performance of Nonnative Speakers of English on TOEFL and Verbal Aptitude Tests.** Angelis, Swinton, and Cowell. October 1979. Gives the results of a study in which 400 graduate and undergraduate applicants took the TOEFL test and either the GRE® verbal or the SAT® verbal and the Test of Standard Written English; includes comparative data on performance across tests.

RR–4. **An Exploration of Speaking Proficiency Measures in the TOEFL Context.** Clark and Swinton. October 1979. Describes a three-year study involving the development and experimental administration of test formats and item types aimed at measuring the English-speaking proficiency of nonnative speakers; results grouped into a prototype Test of Spoken English.

RR–5. **The Relationship Between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language.** Powers. December 1980. Analyzes performance of 6,000 nonnative speakers of English on the TOEFL and GMAT tests; provides support of the basic differences in the two tests and indicates expected GMAT scores for examinees with differing levels of English language proficiency.

RR–6. **Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.** Swinton and Powers. December 1980. Provides evidence that three major factors underlie performance on the TOEFL test; suggests these factors may be interpreted differently for several language groups.

RR–7. **The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings.** Clark and Swinton. December 1980. Examines performance of teaching assistants on the TSE test in relation to their classroom performance as judged by students; finds the TSE test is valid predictor of oral language proficiency for nonnative English-speaking graduate teaching assistants.

RR–8. **Effects of Item Disclosure on TOEFL Performance.** Hale, Angelis, and Thibodeau. December 1980. Assesses the effects of test disclosure by examining the performance on the TOEFL test when a subset of items has been studied prior to an administration; provides separate results by language group and by item type.

RR–9. **Item Performance Across Native Language Groups on the Test of English as a Foreign Language.** Alderman and Holland. August 1981. Examines the performance of different native language groups on TOEFL items; discusses implications for the interpretation and examination of item performance by groups.

RR–10. **Language Proficiency as a Moderator Variable in Testing Academic Aptitude.** Alderman. November 1981. Demonstrates role of language proficiency as moderator variable in assessing academic aptitude; a moderately strong correlation develops between verbal aptitude tests in native and second languages when TOEFL scores indicate high second-language proficiency.

RR–11. **A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979.** Wilson. September 1982. Provides detailed comparative information about personal characteristics, academic aspirations, and test scores of TOEFL examinees by region, native country, and native language.

RR–12. **GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL.** Wilson. October 1982. Analyzes the performance of examinees taking the TOEFL test and either the GRE Aptitude Test or GMAT test; provides further documentation of the relationship between English language proficiency and aptitude test scores earned by foreign students.

RR–13. **The Test of Spoken English as a Measure of Communicative Ability in the Health Professions.** Powers and Stansfield. January 1983. Provides results of using a set of procedures for determining standards of language proficiency in testing pharmacists, physicians, veterinarians, and nurses and for validating the use of the TSE test in health-related professions.

RR–14. **A Manual for Assessing Language Growth in Instructional Settings.** Swinton. February 1983. Describes a methodology for determining the true gains in proficiency that can be expected for students who enter English language training programs at different TOEFL score levels; discusses how the relationship between gains and time enrolled in a program can be used to advise students.

RR–15. **Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students.** Bridgeman and Carlson. September 1983. Describes a survey of faculty in 190 departments at 34 U.S. and Canadian universities with high international student enrollments; respondents indicated a desire to use scores on a direct writing sample to supplement admissions and placement decisions.

RR–16. **Summaries of Studies Involving the Test of English as a Foreign Language, 1963-1982.** Hale, Stansfield, and Duran. February 1984. Includes approximately 80 summaries of empirical research studies involving the TOEFL test, as well as descriptive papers that provide a perspective on history and development of the test.

RR–17. **TOEFL From a Communicative Viewpoint on Language Proficiency: A Working Paper.** Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro. February 1985. Examines the content characteristics of the TOEFL test from a communicative perspective based on current research in applied linguistics and language proficiency assessment.

RR–18. **A Preliminary Study of Raters for the Test of Spoken English.** Bejar. February 1985. Examines the scoring patterns of different TSE raters in an effort to develop a method for predicting disagreements; reports that the raters varied in the severity of their ratings but agreed substantially on the ordering of examinees.

RR–19. **Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English.** Carlson, Bridgeman, Camp, and Waanders. August 1985. Investigates the relationship between essay writing skills and scores on the TOEFL test and the GRE General Test obtained from applicants to U.S. institutions.

RR–20. **A Survey of Academic Demands Related to Listening Skills.** Powers. December 1985. Reports findings from a survey of faculty perceptions regarding the importance of various listening problems of nonnative English-speaking students.

RR–21. **Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference.** Stansfield. May 1986. Includes invited papers and summaries of the discussions that took place at a conference devoted to the TOEFL program's testing of communicative competence.

RR–22. **Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language.** Wilson. January 1987. Analyzes patterns of test taking and score change for examinees who repeated the TOEFL test within 24 to 60 months after they first took the test; shows that repeaters registered substantial average net gains in performance, and differences were noted among national-linguistic groups.

RR–23. **Development of Cloze-Elide Tests of English as a Second Language.** Manning. April 1987. Reports on a study to investigate the validity of cloze-elide tests of English proficiency for students similar to the TOEFL candidate population; suggests that cloze-elide tests are good, indirect measures of English language proficiency , comparing very favorably with more commonly used testing procedures.

RR–24. **A Study of the Effects of Item Option Rearrangement on the Listening Comprehension Section of the Test of English as a Foreign Language.** Golub-Smith. August 1987. Provides evidence that scrambling a test question's answer choices produces differences in both the estimated response functions and equating functions.

RR–25. **The Interaction of Student Major-Field Group and Text Content in TOEFL Reading Comprehension.** Hale. January 1988. Examines the interaction of a student's major-field group with the text content in determining performance on TOEFL reading passages.

RR–26. **Multiple-Choice Cloze Items and the Test of English as a Foreign Language.** Hale, Stansfield, Rock, Hicks, Butler, and Oller. March 1988. Investigates the degree to which multiple-choice cloze items tap reading comprehension, as defined by sensitivity to long-range textual constraints, and tap knowledge of grammar or vocabulary.

RR–27. **Native Language, English Proficiency, and the Structure of the Test of English as a Foreign Language.** Oltman, Stricker, and Barrows. July 1988. Assesses the interrelations among TOEFL items for groups of TOEFL examinees varying in native language and level of English proficiency; concludes that TOEFL construct validity is supported, the test's interpretation varies with examinees' English proficiency, easy and difficult items differ in their potential for diagnosis and global screening, and the dimensionality of the TOEFL test and of competence in English depend on examinees' English proficiency.

RR–28. **Latent Structure Analysis of the Test of English as a Foreign Language.** Boldt. November 1988. Uses IRT-based methods for TOEFL equating; reports a single factor (group) gave a very accurate accounting for the proportions of joint item success.

RR–29. **Context Bias in the Test of English as a Foreign Language.** Angoff. January 1989. Uses a Mantel-Haenszel analysis to test the hypothesis that TOEFL examinees tested

in their native countries are disadvantaged because of American references in the test; concludes that the TOEFL test does not place foreign-tested examinees at a disadvantage.

RR–30. **Accounting for Random Responding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language.** Secolsky. January 1989. Uses two exploratory approaches to determine whether the TOEFL test is speeded according to established criteria; suggests that Section 3 pretest administrations may be slightly speeded, but that further confirmation is needed because of the exploratory nature of the methods.

RR–31. **The TOEFL Computerized Placement Test: Adaptive Conventional Measurement.** Hicks. January 1989. Reports on the development of an experimental TOEFL computerized placement test using conventional scoring methods based on a testing algorithm that routed examinees through item blocks or testlets and permitted backtracking to review answers and change them.

RR–32. **Confirmatory Factor Analysis of the Test of English as a Foreign Language.** Hale, Rock, and Jirele. December 1989. Provides evidence that two major factors underlie performance on the TOEFL test for both low- and high-proficiency examinees of several language groups; helps explain differences in results of earlier factor-analytic research on the TOEFL test.

RR–33. **A Study of the Effects of Variation of Short-term Memory Load, Reading Response Length, and Processing Hierarchy on TOEFL Listening Comprehension Item Performance.** Henning. February 1991. Examines TOEFL listening comprehension item functioning under a variety of controlled stimulus and response conditions; results support a reduction in length of multiple-choice response options for listening comprehension items.

RR–34. **Note Taking and Listening Comprehension on the Test of English as a Foreign Language.** Hale and Courtney. February 1991. Examines effects of note taking in the TOEFL listening comprehension subsection containing short monologues or "minitalks," concludes notetaking produces little benefit in the context of the TOEFL minitalks as they are currently structured.

RR–35. **A Study of the Effects of Contextualization and Familiarization on Responses to the TOEFL Vocabulary Test Items.** Henning. February 1991. Investigates comparative functioning of eight multiple-choice vocabulary item formats; comparative estimates of item difficulty, item discriminability, criterion-related validity, and subtest reliability support the use of vocabulary embedded in reading passages and the use of vocabulary stems with inference-generating information.

RR–36. **A Preliminary Study of the Nature of Communicative Competence.** Henning and Cascallar. February 1992. Provides information on the comparative contributions of some theory-based communicative competence variables to domains of linguistic, discourse, sociolinguistic, and strategic competencies and investigates these competency domains for their relation to components of language proficiencies assessed by the TOEFL, TWE, and TSE tests.

RR–37. **An Investigation of the Appropriateness of the TOEFL Test as a Matching Variable to Equate TWE Topics.** DeMauro. May 1992. Explores the feasibility of using linear and equipercentile equating methods to equate forms of the TWE test using the TOEFL test as an anchor; results suggest the TOEFL and TWE tests do not measure the same skills and examinee groups are often dissimilar in skills.

RR–38. **Scalar Analysis of the Test of Written English.** Henning. August 1992. Investigates the psychometric characteristics of the TWE rating scale using Rasch model scalar analysis; results suggest the intervals between the TWE scale steps are uniform and the size of the intervals is appropriately larger than the error associated with assignment of individual ratings.

RR–39. **Effects of the Amount of Time Allowed on the Test of Written English.** Hale. June 1992. Examines student performance on the TWE test under two time limits—30 and 45 minutes; results indicated mean scores were higher by 1/4 to 1/3 point under the 45-minute condition, but additional time had little effect on the relative standing of students on the test.

RR–40. **Reliability of the Test of Spoken English Revisited.** Boldt. November 1992. Examines effects of scale, section, examinee, and rater as well as the interactions of these factors on the TSE test; offers suggestions for improving reliability.

RR–41. **Distributions of ACTFL Ratings by TOEFL Score Ranges.** Boldt, Larsen-Freeman, Reed, and Courtney. November 1992. Examines cross-tabulations of students' TOEFL section scores with listening, reading, and writing proficiency rated according to ACTFL Proficiency Guidelines descriptors; provides distributions of ACTFL ratings for levels of TOEFL section scores that may be helpful in interpreting TOEFL scores in terms of language performance.

RR–42. **Topic and Topic Type Comparability on the Test of Written English.** Golub-Smith, Reese, and Steinhaus. March 1993. Analyzes scores obtained on eight prompts (differing in both subject matter and level of explicitness with which the essay task was presented) spiraled worldwide at the October 1989 TWE administration; results suggest that although differences among prompts were small, further investigation of differences observed at some score levels is warranted.

RR–43. **Uses of the Secondary Level English Proficiency (SLEP) Test: A Survey of Current Practice.** Wilson. March 1993. Provides information regarding testing practices and purposes, characteristics of examinees, test users' perceptions of the principal strengths and limitations of the test and test manual, and the extent and nature of local studies concerned with validating the SLEP test.

RR–44. **The Prediction of TOEFL Reading Comprehension Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items.** Freedle and Kostin. May 1993. Explores predictors of reading comprehension item difficulty and compares influence of item difficulty at five different verbal ability levels; concludes that a significant amount of item difficulty variance can be accounted for by a relatively small number of variables for the three reading item types studied.

RR–45. **Test-Retest Analyses of the Test of English as a Foreign Language.** Henning. June 1993. Provides comparative global and component estimates of TOEFL test-retest, alternate forms, and internal consistency reliability as well as information about differential change in subtest difficulty on repeated application over a small interval of time; study was limited by small sample size.

RR–46. **Multimethod Construct Validation of the Test of Spoken English.** Boldt and Oltman. December 1993. Uses factor analysis and multidimensional scaling to explore the relationships among TSE subsections and rating dimensions; results show the roles of test section and proficiency scales in determining TSE score variation.

RR–47. **An Investigation of Proposed Revisions to Section 3 of the TOEFL Test.** Schedl, Thomas, and Way. March 1995. Examines speededness of a prototype revised TOEFL in which discrete vocabulary items have been replaced by additional reading comprehension questions; results support use of five reading passages with a total of 50 questions and suggest that no less than 55 minutes testing time be allowed.

RR–48. **Analysis of Proposed Revisions of the Test of Spoken English.** Henning, Schedl, and Suomi. March 1995. Compares a prototype revised TSE with the original version of the test with respect to interrater reliability, frequency of rater discrepancy, component task adequacy, scoring efficacy, and other aspects of validity; results underscore the psychometric quality of the revised TSE.

RR–49. **A Study of the Characteristics of the SPEAK Test.** Sarwark, Smith, MacCallum, and Cascallar. March 1995. Investigates issues of reliability and validity associated with the original locally administered and scored SPEAK test, the "off-the-shelf" version of the original TSE; results indicate that this version of the SPEAK test is reasonably reliable for local screening and is an appropriate measure of English-speaking proficiency in U.S. instructional settings.

RR–50. **A Comparison of Performance of Graduate and Undergraduate School Applicants on the Test of Written English.** Zwick and Thayer. May 1995. Compares undergraduate and graduate students matched on TOEFL total score; the matched undergraduate students had higher scores on the TWE test, a different result than comparisons based on unmatched groups.

RR–51. **An Analysis of Factors Affecting the Difficulty of Dialog Items in TOEFL Listening Comprehension.** Nissan, DeVincenzi, and Tang. February 1996. Identifies five features of TOEFL dialogue items that were significantly related to item difficulty.

RR–52. **Reader Calibration and Its Potential Role in Equating for the Test of Written English.** Myford, Marr, and Linacre. May 1996. Uses FACETS, a Rasch-based procedure, to calibrate TWE readers and provides information on reader characteristics and their influence on rating, and whether readers can be treated as interchangeable.

RR–53. **An Analysis of the Dimensionality of TOEFL Reading Comprehension Items.** Schedl, Gordon, Carey, and Tang. March 1996. Investigates the dimensionality of the TOEFL reading test; confirmatory analyses did not support a separate "reasoning" factor among the reading items, but exploratory analyses indicated the possibility of a second factor related to passage content or position.

RR–54. **A Study of Writing Tasks Assigned in Academic Degree Programs.** Hale, Taylor, Bridgeman, Carson, Kroll, and Kantor. June 1996. Develops a classification scheme for undergraduate and graduate writing tasks across a sample of disciplines and institutions; different types of writing assignments were characterized, and differences among disciplines in writing demands were examined.

RR–55. **Adjustment for Reader Rating Behavior in the Test of Written English.** Longford. August 1996. Evaluates the potential impact of one method for adjustment of TWE scores due to rater differences; the method can reduce error in TWE scores and could be used to combine information across rating exercises to further increase measurement precision.

RR–56. **The Prediction of TOEFL Listening Comprehension Item Difficulty for Minitalk Passages: Implications for Construct Validity.** Freedle and Kostin. August 1996. Relevant features of item passages significantly influenced listening comprehension item difficulty, indicating that listeners were responding to the meanings of the passages.

RR–57. **Survey of Standards for Foreign Student Applicants.** Boldt and Courtney. August 1997. The survey found that minimum TOEFL scores were usually set by reference to policies of other institutions. Commonly used minimum scores were tabulated. Minimums were usually used to route student into further English training, and not to reject applicants.

RR–58. **Using Just Noticeable Differences to Interpret Test of Spoken English Scores.** Stricker. August 1997. The study assessed the difference in scores needed before observers discern a difference in English proficiency of international teaching assistants. The Just Noticeable Differences estimates appeared to be useful for interpreting the practical significance of TSE scores.

RR–59. **Computer Familiarity Among TOEFL Examinees.** Kirsch, Jamieson, Taylor, and Eignor. March 1998. This report profiles approximately 90,000 TOEFL examinees in terms of their access to and experience with computers. Overall, some 16% of the TOEFL population was judged to have low computer familiarity, another 34% to have moderate familiarity, and approximately 50% to have high familiarity. The report also examines computer familiarity in terms of a number of examinee background characteristics.

RR–60. **Development of a Scale for Assessing the Level of Computer Familiarity of TOEFL Examinees.** Eignor, Taylor, Kirsch, and Jamieson. March 1998. This paper describes in greater detail the development of the scale used to profile TOEFL examinees in terms of their computer familiarity (see TOEFL RR-62). It also details the procedures used to assess the underlying factor structure of the complete questionnaire.

RR–61. **The Relationship Between Computer Familiarity and Performance on Computer-based TOEFL Test Tasks.** Taylor, Jamieson, Eignor, and Kirsch. March 1998. This paper reports on the effects of computer familiarity for a group of low- and high-computer familiar TOEFL examinees' performance on a set of 60 computer-based TOEFL tasks. This report concludes that, after administration of a

computer tutorial, and controlling for language ability, no evidence of adverse effects on TOEFL CBT performance were found due to lack of prior computer experience.

RR–62. **Designing and Evaluating a Computer-Based TOEFL Tutorial.** Jamieson, Taylor, Kirsch, and Eignor. March 1999. This report describes the development of a computer-based TOEFL tutorial and the experiences of the 1,169 individuals who participated in a computer familiarity study. These analyses took into account both computer familiarity and English ability, which proved to be important in explaining some differences in time to complete the tutorials and perception of the tutorials' usefulness. As a result of the study, some changes were made before operational implementation of the computerized TOEFL test to reduce the time needed to complete the tutorials.

RR–63. **Validating the Revised Test of Spoken English Against a Criterion of Communicative Success.** Powers, Schedl, Wilson-Leung, and Butler. March 1999. A communicative competence orientation was taken to study the validity of test score inferences derived from the TSE test. Student evaluations were captured by devising and administering a secondary listening test (SLT) to assess students' understanding of Test of Spoken English examinees' speech, as represented by their taped responses to tasks on the TSE test. The objective was to determine the degree to which official TSE scores are predictive of listeners' ability to understand the messages conveyed by TSE examinees.

RR–64. **Computer Analysis of the TOEFL Test of Written English.** Frase, Faletti, Ginther, and Grant. May 1999. Test of Written English essays from several language groups were analyzed, with essays from English-speaking examinees providing a baseline for comparison with the essays by examinees with English as a second language (ESL). Analyses revealed topic differences influenced some essay variables, but language groups were not differentially affected; language groups appeared to differ in the extent of directness, expressiveness, and academic stance of their writing styles, and number of words and the average length of words taken together are predictive of TWE essay scores of ESL writers.

RR–65. **Monitoring Sources of Variability Within the Test of Spoken English Assessment System.** Myford and Wolfe. June 2000. The purposes of this study were to examine four sources of variability within the Test of Spoken English assessment system, to quantify ranges of variability for each source, to determine the extent to which these sources affect examinee performance, and to highlight aspects of the assessment system that might suggest a need for change. Data obtained from the February and April 1997 TSE scoring sessions were analyzed using *Facets*.

RR–66. **Effects of the Presence and Absence of Visuals on Subjects' Performance on TOEFL CBT Listening Comprehension Stimuli.** Ginther. August 2001. This study investigated the effects of different types of visual presentations, looking specifically at the effects of language proficiency (high or low), still photos (present or absent), and type of stimuli (dialogues/short conversations, academic discussions, minitalks with context visuals, minitalks with content visuals) on performance on standard multiple-choice listening items. Three two-way interactions were significant: proficiency by type of stimuli, type of stimuli by time, and type of stimuli by visual condition. The majority of the subjects indicated a strong preference for the presence of visuals.

RR–67. **Automatic Assessment of Vocabulary Usage Without Negative Evidence.** Leacock and Chodorow. November 2001. As part of the TOEFL program's effort to develop performance- based measures of communicative competence, we developed an automated statistical method for assessing an examinee's use of vocabulary words in constructed responses. Our error-detection system, ALEK (Assessing Lexical Knowledge), infers negative evidence from the low frequency or absence of constructions in 30 million words of well-formed, copy-edited text from North American newspapers. The system evaluated word usage in essay length responses to TOEFL prompts. The system performed with about 80% precision and 20% recall.

RR–68. **Influence of Irrelevant Speech on Standardized Test Performance.** Powers, Albertson, Florek, Malak, Johnson, Nemceff, Porzuc, Silvester, Wang, Weston, Winner, and Zelazny. Winter 2002. The aim of this study was to estimate the impact of distraction on test performance and evaluate ways to reduce it. The distraction of interest was from fellow examinees taking a speaking test. Attempts to reduce distraction to an acceptable level were largely unsuccessful. Impact on actual test performance, however, was slight to negligible Intermingling examinees with others who are taking a speaking test remains a concern, primarily because of strong negative perceptions by test takers.

RR–69. **The Performance of Native Speakers of English and ESL Speakers on the TOEFL-CBT and GRE General Test.** Stricker. Spring 2003. The study examined construct validity of the TOEFL-CBT. Two samples of GRE test takers were used: native speakers of English specially recruited to take the TOEFL-CBT, and ESL test takers who routinely took the TOEFL-CBT recently. Native speakers performed well on TOEFL, relative to ESL test takers and to the maximum possible scores on the test, and varied less in their test performance than did ESL test takers. All of the findings are consistent with previous results with the paper-and-pencil TOEFL, support the construct validity of the TOEFL-CBT, and illuminate its interplay with ability tests for ESL test takers.

RR–70. **Exploring Variability in Judging Writing Ability in a Second Language: A Study of Four Experienced Raters of ESL Compositions.** Erdosy. Spring 2004. In this study, four raters constructed scoring criteria while assessing corpora of 60 TOEFL essays without the aid of a scoring rubric. The study revealed key points in the decision-making process, where raters' behavior diverged, and examined the impact of prior experience on these. The identification of such divergences, and potential explanations for them lay the foundations for a principled explanation of rater variability.

RR–71. **Investigating the Validity of TOEFL: A Feasibility Study Using Content and Criterion-Related Strategies.** Rosenfeld, Oltman, and Sheppard. Spring 2004. This study investigated the feasibility of two complementary approaches to assess the validity of the TOEFL: evidence based on test content and a criterion-related validation strategy. The former approach involved item-rating procedures to evaluate and document the relationship between the language tasks or

behaviors previously identified as important for academic success and the test items used to measure them. In the second approach, experimental rating scales were developed for use by faculty to evaluate students' current level of English language proficiency. These scales were designed to sample the domain of behaviors previously identified as important.

RR–72. **An Investigation of the Impact of Composition Medium on the Quality of Scores from the TOEFL Writing Section: A Report from the Broad-Based Study.** Wolfe and Manolo. March 2004. This study examined scores from 133,906 operationally scored TOEFL essays. Results demonstrate that scores assigned to word-processed essays are slightly more reliable and exhibit higher correlations with scores from the TOEFL multiple-choice sections. In addition, a main effect exists for gender; age, continent, native language, and English proficiency exhibit interactions in their relationships with composition-medium choice. Finally, although there were no differences observed between handwritten and word-processed essay scores, when differences in overall English proficiency between composition medium groups are controlled, a large interaction emerges.

RR–73. **Beyond Essay Length: Evaluating e-rater's Performance on TOEFL Essays.** Chodorow and Burstein. February 2004. This study examines the relation between essay length and holistic scores assigned to TOEFL essays by e-rater. Results show that an early version of the system, e-rater99, accounted for little variance in human reader scores beyond that which could be predicted by essay length. A later version of the system, e-rater01, performs significantly better than its predecessor and is less dependent on length due to its greater reliance on measures of topical content and of complexity and diversity of vocabulary.

RR–74. **Elicited Speech from Graph Items on the Test of Spoken English.** Katz, Xi, Kim, and Cheng. February 2004. This research applied a cognitive model to identify item features that lead to irrelevant variance on the TSE. The TSE includes an item that elicits a description of a statistical graph. We adapted a cognitive theory of graph comprehension to predict the degree to which TSE graph items tap irrelevant skills such as graph reading. Through analyses of existing TSE data as well as an experiment, we show how the theory provides specific, empirically justified recommendations on the construction of graph items that minimize the influence of extraneous skills.

RR–75. **Comparability of TOEFL CBT Writing Prompts: Response Mode Analyses.** Breland, Lee, and Muraki. July 2004. Eighty-three TOEFL-CBT writing prompts administered between July 1998 and August 2000 were examined for differences attributable to the response mode (handwritten or word-processed) chosen by examinees. Although there was little observed difference in mean writing scores, when examinees were matched on English language ability, small differences were observed in effect sizes consistently favoring the handwritten-response mode. The difference favoring the handwritten-response mode occurred for all of the writing prompts analyzed, which suggests a general effect for response mode. The differences for individual writing prompts were small, however.

RR-76. **An Analysis of TOEFL-CBT Writing Prompt Difficulty and Comparability for Different Gender Groups.** Breland, Lee, Najarian, and Muraki. February 2004. Comparability of writing assessment prompts was investigated in two phases. In Phase I, 47 writing prompts administered in the TOEFL-CBT July–December 1998 were examined. Logistic regression procedures estimated prompt difficulty and gender effects. A taxonomy of prompt characteristics was developed and related to prompt difficulty and gender differences. In Phase II, 87 prompts administered from July 1998–March 2000 were analyzed. All of the prompts in Phase I with 40 new prompts were analyzed using the larger Phase II database. Recommendations are made for statistical quality control procedures to identify less comparable prompts.

RR–77. **Comparability of TOEFL-CBT Writing Prompts for Different Native Language Groups.** Lee, Breland, and Muraki. August 2004. This study investigated the comparability of TOEFL-CBT writing prompts for examinees of different native language backgrounds. Eighty-one prompts introduced July 1998–August of 2000 were examined using a three-step logistic regression procedure for ordinal items. An English language ability variable was created by summing the standardized TOEFL reading, listening, and structure scale scores and used to match examinees of East Asian (Chinese, Japanese, and Korean) and European (German, French, and Spanish) language groups. Although about one third of the 81 prompts were initially flagged because of statistically significant group effects, the effect sizes were too small to be classified as having an important group effect.

RR-78. **Toward Accessible Computer-Based Tests: Prototypes for Visual and Other Disabilities.** Hansen, Forer, and Lee. November 2004. To ensure that computer-based tests are accessible to individuals with disabilities, three prototype test delivery systems were developed: (a) the Self-Voicing Test version 3 (SVT3), (b) the HTML-Form System (HFS), and (c) the Visually Oriented System (VOS). SVT3 provided built-in text-to-speech capabilities and keyboard operation. HFS used standard HTML form input elements and supported the optional use of text-to-speech and Braille. VOS was visually oriented and operable via mouse. Fifteen adults, two to four from each of the six disability statuses—blindness, low vision, deafness, deaf-blindness, learning disability, and no disability—evaluated the systems. Items came from the domains of reading comprehension, listening comprehension, structure (grammar), writing, and math. The study found that although all the systems had weaknesses, 13 of the participants recommended at least one of the delivery methods.

RR-79. **Exploring Item Characteristics That Are Related to the Difficulty of TOEFL Dialogue Items.** Kostin. July 2004. This study explored the relationship between a set of item characteristics and the difficulty of TOEFL dialogue items. Identifying characteristics related to item difficulty may improve the efficiency of the item-writing process. The study employed 365 TOEFL dialogue items coded on 49 variables. Three of the five significant variables in Nissan, DeVincenzi, and Tang correlated with item difficulty in this study. Eleven variables met a critical probability criterion. Multiple-regression analyses indicate that the variables account for about 40% of the variance in item difficulty.

Technical Reports

TR–1. **Developing Homogeneous Scales by Multidimensional Scaling.** Oltman and Stricker. February 1991. Explores feasibility and value of using cluster scores; reports that corresponding scores for clusters and test sections do not differ in internal-consistency reliabilities and intercorrelations for total sample, but diverge inconsistently for high-scoring and low-scoring examinees.

TR–2. **An Investigation of the Use of Simplified IRT Models for Scaling and Equating the TOEFL Test.** Way and Reese. February 1991. Explores the use of two alternative IRT estimation models for scaling and equating the TOEFL test; results support the use of the three-parameter model.

TR–3. **Development of Procedures for Resolving Irregularities in the Administration of the Listening Comprehension Section of the TOEFL Test.** Way and McKinley. February 1991. Evaluates two procedures, an analysis of covariance and a Bayesian procedure, for determining whether examinees in a given test center are affected by a testing irregularity on the listening comprehension section of the TOEFL test; recommends the use of both approaches in resolving testing irregularities.

TR–4. **Cross-Validation of a Proportional Item Response Curve Model.** Boldt. April 1991. Investigates whether a proportional item response curve (PIRC) model could serve as a basis for simpler equating methods than are currently used by the TOEFL program; PIRC, a three-parameter logistic model, and a modified Rasch model prediction are approximately equally accurate, and the estimation sample size seems to make little difference.

TR–5. **The Feasibility of Modeling Secondary TOEFL Ability Dimensions Using Multidimensional IRT Models.** McKinley and Way. February 1992. Explores the feasibility of an IRT-based method of modeling examinee performance on secondary ability dimensions of the TOEFL test; results indicate multidimensional IRT and confirmatory multidimensional IRT models provide corroborative evidence in interpreting the structure of the test.

TR–6. **An Exploratory Study of Characteristics Related to IRT Item Parameter Invariance with the Test of English as a Foreign Language.** Way, Carey, and Golub-Smith. September 1992. Explores features of TOEFL test items that may contribute to a lack of IRT item parameter invariance; results suggest several possible factors that may contribute to a lack of IRT item parameter invariance, and researchers offer suggestions to improve the IRT item parameter invariance of TOEFL test items.

TR–7. **The Effect of Small Calibration Sample Sizes on TOEFL IRT-Based Equating.** Tang, Way, and Carey. December 1993. Compares performance of LOGIST and BILOG on TOEFL IRT-based scaling and equating, using both real and simulated data and two calibration structures; results suggest retaining pretest sample sizes of 1,000 for LOGIST if possible.

TR–8. **Simulated Equating Using Several Item Response Curves.** Boldt. January 1994. Examines several item response models as bases for TOEFL equating, using simulation trials to equate the test to itself; for variations of sample size and anchor test difficulty, reports on discrepancies between scores identified as comparable.

TR–9. **Investigation of IRT-Based Assembly of the TOEFL Test.** Chyn, Tang, and Way. March 1995. Investigates the feasibility of the Automated Item Selection (AIS) procedure for the TOEFL test, using statistical specifications based on item response theory (IRT); results suggest that Record Examinations and language proficiency testing AIS-assembled TOEFL tests have greater statistical consistency than tests assembled by traditional means and can successfully meet the IRT-based specifications.

TR–10. **Estimating the Effects of Test Length and Test Time on Parameter Estimation Using the HYBRID Model.** Yamamoto. March 1995. Examines speededness of a prototype TOEFL reading comprehension section with a model that determines when each examinee switches from an ability-based response strategy to a strategy of responding randomly; results show that proportionately more examinees were affected by test speededness when given a 50-minute time limit than when given a 55- or 60-minute time limit, with little difference between 55- and 60-minute limits.

TR–11. **Using a Neural Net to Predict Item Difficulty.** Boldt and Freedle. November 1996. Uses a neural net approach to add nonlinear prediction to linear methods of predicting item difficulty. Several variables added by the neural net improved prediction. However, substantial capitalization on chance can occur in this type of study, which can weaken substantive inferences.

TR–12. **How Reliable Is the TOEFL Test?** Wainer and Lukehele. August 1997. Examines overestimation of the reliability of reading comprehension testlets due to local dependence when various test sections are analyzed individually; concludes that the test was unidimensional enough for the use of univariate IRT to be efficacious.

TR–13. **Concurrent Calibration of Dichotomously and Polytomously Scored TOEFL Items Using IRT Models.** Tang and Eignor. August 1997. Combines reading and writing, and listening and speaking items to approximate one level of integration that might be adopted in the future. The item combinations could be successfully calibrated using a 3PL model combined with either a generalized partial credit model or a graded response model.

TR–14. **Graphical Models and Computerized Adaptive Testing.** Almond and Mislevy. March 1998. Synthesizes ideas from graphical modeling and educational testing, pointing out how variables can enter the modeling process in validity studies, task construction, test assembly, response characterization, and in the student model.

TR–15. **Strengthening the Ties that Bind: Improving the Linking Network in Sparsely Connected Rating Designs.** Myford and Wolfe. August 2000. Evaluated effectiveness of linking raters when large numbers are involved in scoring sessions and rater overlap is minimal. Findings indicate efficacy of benchmark sets for establishing at least the minimal connectivity needed in the rating design to allow placement of raters and examinees on a single scale.

TR–16. **Using a New Statistical Model for Testlets to Score TOEFL.** Wainer and Wang. May 2001. Modifies the standard three-parameter IRT model to include an additional random effect for items nested within the same testlet. This parameter characterizes the amount of local dependence in a testlet.

TR–17. **A Study of the Use of Collateral Statistical Information in Attempting to Reduce TOEFL IRT Item Parameter Estimation Sample Sizes.** Tang and Eignor. June 2001. Examined procedures that might allow reduction in pretest sample sizes needed for IRT calibration for item pools used in computer-based testing programs.

Additional TOEFL Monographs

MS–1. **A Review of the Academic Needs of Native English-Speaking College Students in the United States.** Ginther and Grant. September 1996. Surveys literature concerning the academic needs of native English-speaking college students in the United States from several perspectives; concludes with questions about the identification of the appropriate testing domain, the appropriate level of specifications of test tasks, the fairness of testing academic tasks, and authentic language use in testing.

MS–2. **Polytomous Item Response Theory (IRT) Models and Their Applications in Large-Scale Testing Programs: Review of Literature.** Tang. September 1996. Reviews two commonly used polytomous IRT models: the generalized partial credit model and the graded response model. Also reviews programs and procedures for calibrating dichotomously and polytomously scored items and the application of models in large-scale testing programs.

MS–3. **A Review of Psychometric and Consequential Issues Related to Performance Assessment.** Carey. September 1996. Summarizes the psychometric and consequential issues involved in the use of performance assessments that are of relevance to TOEFL 2000; results from performance assessments show that there is a high degree of task-specific variance, that the magnitude of rater variance can be minimized, and that they can be context bound and of limited generalizability.

MS–4. **Assessing Second Language Academic Reading from a Communicative Competence Perspective: Relevance for TOEFL 2000.** Hudson. September 1996. Examines issues involved in the assessment of academic reading from a communicative proficiency perspective; concludes with implications for academic reading assessment, paying particular attention to the four validity components of construct validity, value implications, relevance/utility, and social consequences.

MS–5. **TOEFL 2000–Writing: Composition, Community, and Assessment.** Hamp-Lyons and Kroll. March 1997. Explores the salient issues of an approach of assessing writing in the context of the TOEFL test and in light of what is currently known/believed about the acquisition and assessment of writing; describes various approaches to writing assessment that might be used and considers how these might be applied to academic writing in the TOEFL 2000 context.

MS–11 **Technologies for Language Testing.** Frase, Gong,

MS–6. **A Review of Research Into Needs in English for Academic Purposes of Relevance to the North American Higher Education Context.** Waters. November 1996. Examines research into needs in EAP of relevance to the North American higher education context concludes that there is no existing body of research that could form an adequate basis for the development of a test of EAP and proposes a program of further research.

MS–7. **The Revised Test of Spoken English: Discourse Analysis of Native Speaker and Nonnative Speaker Data.** Lazaraton and Wagner. December 1996. Describes a qualitative discourse analysis of native speaker and nonnative speaker responses to the revised TSE test; results indicated that the match between intended task functions (as per the content specifications) and the actual functions employed by native speakers was quite close.

MS–8. **Testing Speaking Ability in Academic Contexts: Theoretical Considerations.** Douglas. April 1997. This paper provides a theoretical background for the large-scale assessment of speaking ability for undergraduate/graduate university admissions of international students. It argues that speech production and comprehension are systematically integrated, language knowledge is multicomponential, and strategic ability is central to the interpretation of context in the test assessment of speaking ability.

MS–9. **Theoretical Underpinnings of the Test of Spoken English Revision Project.** Douglas and Smith. May 1997. This paper lays out a theoretical foundation for the revision of the Test of Spoken English. It discusses communicative competence, sociolinguistic and discourse factors that influence spoken language performance, test method characteristics that influence performance, as well as types of evidence necessary for establishing reliability and validity of the revised test.

MS–10. **Communicative Language Proficiency: Definition and Implications for TOEFL 2000.** Chapelle, Grabe, and Berns. May 1997. Discussion of TOEFL 2000 in the TOEFL Committee of Examiners' meetings resulted in a framework representing components believed to be relevant in defining language use in an academic context. This paper describes the framework and serves as a record of past discussions that can inform future work on the TOEFL 2000 project.

MS–18. **TOEFL 2000 Writing Framework: A Working**

Hansen, Kaplan, Katz, and Singley. July 1998. This paper reviews current and emerging technologies relevant to language assessment. It addresses the cognitive and social technologies that are needed to support efficient technology based language assessment, reviews hardware, software, and item development technologies, and discusses implications for new test development.

MS–12. **Computer and Communications Technologies in Colleges and Universities in the Year 2000.** Hansen and Willut. March 1998. This report describes the current environment in colleges and universities with respect to computer and communications technologies and examines a number of factors that are necessitating change in that environment. It attempts to anticipate how changes in computer and communications technologies in North American colleges and universities by the year 2000 might change the way in which students do their work.

MS–13. **A Review of Computer-Based Speech Technology for TOEFL 2000.** Burstein, Kaplan, Rohen-Wolff, Zuckerman, and Lu. September 1999. As part of our ongoing effort to examine enabling and important technologies, we review the state of the art in computer-based speech technology in the context of the Test of English as a Foreign Language testing program. The goal is to assess the readiness of various computer-based speech technologies for this testing program. This paper focuses on desktop applications for speech recognition and speech synthesis.

MS–14. **Looking Back, Looking Forward: Trends in Intensive English Program Enrollments.** Powell. April 2001. To create credible forecasts for enrollment trends in intensive English language programs (IEPs) in the United States, we analyzed past influences on IEP enrollments. We reviewed available censuses of international students and related to the circumstances external to English programs that seemed to affect the movement of students into IEPs. Findings suggest that IEP administrators should pay close attention to events occurring outside the English program to anticipate future enrollments and to position their programs to respond to changes in the intensive English market.

MS–15. **Washback in Language Testing.** Bailey. June 1999. This monograph summarizes recent research on language testing washback, poses a model for washback, and examines available research related to this model. Recommendations for appropriate research methods to be used in future investigations of washback are made.

MS–16. **TOEFL 2000 Framework: A Working Paper.** Jamieson, Jones, Kirsch, Mosenthal, and Taylor. April 2000. This paper lays out a preliminary working framework for the development of the TOEFL 2000 test. The goal of framework is to guide development of more specific frameworks and research agendas for assessing reading, writing, listening, and speaking, as both independent and integrated modalities.

MS–17. **TOEFL 2000 Reading Framework: A Working Paper.** Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, and Schedl. April 2000. This monograph on the assessment of reading comprehension addresses the proposed TOEFL 2000 framework described in Jamieson et al. and defines how it can be realized and implemented in a test of reading comprehension.

Paper. Cumming, Kantor, Powers, Santos, and Taylor. April 2000. This monograph on the assessment of writing addresses the proposed TOEFL 2000 framework described in Jamieson et al. and defines how it can be realized and implemented in a test of writing proficiency.

MS–19. **TOEFL 2000 Listening Framework: A Working Paper.** Bejar, Douglas, Jamieson, Nissan, and Turner. September 2000. This monograph is an initial attempt to define listening as it will be measured in the TOEFL 2000 test, within the framework delineated in Jamieson et al.

MS–20. **TOEFL 2000 Speaking Framework: A Working Paper.** Butler, Eignor, Jones, McNamara, and Suomi. June 2000. This monograph on the modality of speaking addresses the proposed TOEFL 2000 framework described in Jamieson et al. It also considers ways in which the new TOEFL 2000 speaking component will improve upon the current version of the TOEFL test and the Test of Spoken English.

MS–21. **The Reading, Writing, Speaking, and Listening Tasks Important for Academic Success at the Undergraduate and Graduate Levels.** Rosenfeld, Leung, and Oltman. November 2001. This project attempted to operationalize the theoretical frameworks for TOEFL into statements of tasks undergraduate and graduate students would need to perform to complete their academic programs.

MS–22. **Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation Into Raters' Decision Making, and Development of a Preliminary Analytic Framework.** Cumming, Kantor, and Powers. December 2001. This project established a framework to describe the decision-making processes that experienced writing assessors use to evaluate ESL written compositions. The framework will assist in the development and field testing of a scoring scheme for the writing component of a new TOEFL.

MS–23. **The Effects of Notetaking, Lecture Length, and Topic on the Listening Component of TOEFL 2000.** Carrell, Dunkel, and Mollaun. Winter 2002. The study examined effects of notetaking, lecture length, and topic, as well as two aptitude variables, on listening comprehension with ESL students representative of the TOEFL population. Results revealed positive effects for notetaking and lecture length, as well as significant interactions between notetaking and topic and between notetaking and lecture length. No differences in the pattern of results occurred when listening comprehension proficiency and short-term memory were taken into consideration along with the three main factors.

MS–25. **Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus.** Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay, and Urzua. Spring 2004. Because there have been few large-scale empirical investigations of academic registers and virtually no investigations of spoken academic registers, it has been almost impossible to evaluate the representativeness of ESL/EFL materials and assessment instruments. The TOEFL 2000 spoken and written academic language (T2K-SWAL) corpus was constructed and analyzed to help fill this gap. This report describes the design and analysis of the corpus.