

The Researcher

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

Spring 2003

Automatic Essay Scoring: A Tool for Learning

The direct assessment of academic writing skills is highly valued by teachers of English as a second language and college and university faculty and administrators. Both the Test of Written English™ (TWE®) and TOEFL® CBT recognize this and have test takers respond to a brief question or topic by writing an essay, which two highly trained human raters score. The new TOEFL under development also will include the direct assessment of writing.

ESL teachers support the inclusion of direct writing tasks in assessments, believing that they encourage students to practice writing. The best way to learn to write is *to write and to receive feedback* about the quality of this writing. However, opportunities for practice and feedback are limited by the availability of teachers to provide such feedback. This is where automated essay scoring can provide support for learning.

A Web-based service, *ScoreItNow!*™, uses an automated essay scoring system, *e-rater*®, to provide test takers an opportunity to practice writing TOEFL essays and to receive feedback on the quality of their performance. This product illustrates how new technologies, after careful research and development, can support ESL learning.

E-rater, developed at ETS, uses Natural Language Processing (NLP) tools and regression analyses to model the scoring systems of human raters. To develop the model, a sample of essays on a specific topic is scored by human raters. Next, NLP tools are used to analyze more than 50 syntactic, discourse, topical, and lexical features of a set of these scored essays. Then step-wise regression is used to construct a model of the NLP features that predict the scores of human raters for a subset of the scored essays. This model then generates *e-rater* scores for the remaining essays the human raters have scored. The accuracy of the *e-rater* model is evaluated by comparing the *e-rater* scores with those of the human raters. *E-rater* also detects atypical essays and issues advisory warnings about such essays.

E-rater is used as a second rater for essays on the Graduate Management Admission Test®. While not used operationally for TOEFL CBT or TWE, a number of researchers have conducted investigations using *e-rater* to score TOEFL essays, and these efforts have led to improvements in *e-rater*. Burstein and Chodorow (1999) used an early version of *e-rater* to score a sample of TOEFL essays that had been previously scored by

human raters. Although Burstein and Chodorow reported very high agreement between scores assigned to essays by *e-rater* and by human raters, they also noted some scoring anomalies for different language groups. On the average, *e-rater* assigned scores that were slightly higher than human raters for essays written by native speakers of some languages and those of slightly lower than those of human raters for speakers of other languages. Sheehan (in preparation) investigated possible explanations for these discrepancies. She found that although scores of both *e-rater* and human raters were related to essay length, this relationship was stronger for *e-rater* than for human raters. Although this early version of *e-rater* did not directly measure essay length, some of the features used in the scoring models were highly correlated with essay length. Furthermore, Sheehan argues that the discrepancies in *e-rater* and human scores for different language groups could be explained by group differences in essay length characteristics.

Subsequently, a new version of *e-rater* was developed. *E-rater* was enhanced through improved syntactic analysis and the addition of such text (lexical) complexity features as the number of unique words and the number of longer words. In a recent study, Chodorow and Burstein (2002) compared the performance of this new version with that of both the earlier version of *e-rater* and the human raters. For the TOEFL essays examined in this study, the exact agreement rates were .46 between the earlier version of *e-rater* and the human rating, .53 between the new version of *e-rater* and the human rating, and .56 between two human raters. This study is a clear example of research contributing to product improvement.

ScoreItNow! uses this newer version of *e-rater*, and the expectation is for even more improvement in automated essay scoring in the future as new features of yet unmeasured qualities of writing are identified and incorporated into the system.

References

- Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In M. Olsen (ed.), *Proceedings of the Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. (pp 68-75). Joint symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies. College Park: University of Maryland.
- Chodorow, M., & Burstein, J. (2002). Beyond essay length: Evaluating *e-rater's* performance on TOEFL essays. Manuscript submitted for publication.
- Sheehan, K. M. (2002). Discrepancies in human and computer-generated essay scores for TOEFL-CBT essays. Manuscript submitted for publication.



TOEFL Research and Technical 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, 56, 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, 46, 48, MS7, MS9 | 19, 36, 38, 64 | MS1, MS4, MS5, MS6, MS8, TR14, MS18, MS19, MS20 |
| Face/Content Validity | 1, 16, 17, 21 | 1, 2, 16, 17, 20, 21, 71 | 1, 16, 17, 21, 71 | 1, 16, 17, 21, 71 | 49 | 15, 19, 39, 54 | MS10, MS21, MS25 |
| 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 | |
| Response Validity | | | | | | 38 | |
| TEST INFORMATION | | | | | | | |
| Score Interpretation | 3, 5, 10, 12, 36, 41, TR11 | 3, 5, 10, 12, 36, 41, TR11 | 3, 5, 10, 12, 36, 41, TR11 | 3, 5, 10, 12, 36, 41, TR11 | 36 | 36, 38 | MS3 |
| Underlying Processes | 36 | 33, 36 | 36 | 36 | 36 | 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 | 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 | 9, 25, 29 | | | 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 | 37, 38, 42, 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, TR15 | 38 | MS22, MS24, 70 |
| 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 |
| Alternate Forms | 45 | 45 | 45 | 45 | | 42 | |
| Test-Retest | 45, TR6 | 45, TR6 | 45, TR6 | 45, TR6 | | | |
| Inter-/Intrater | | | | | 4, 7, 18, 40, 49, TR15 | 19, 38, 55 | |
| APPLIED TECHNOLOGY | | | | | | | |
| Innovative Formats | 2, 23, 61, 62 | 2, 23, 26, 33, 34, 66 | 2, 23, 26 | 2, 23, 26, 35 | | | 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."

New Research Reports

RR – 69. The Performance of Native Speakers of English and ESL Speakers on the TOEFL-CBT and GRE General Test

Lawrence J. Stricker
Spring 2003

The purpose of this study was to replicate previous research on the construct validity of the paper-and-pencil version of the TOEFL test and extend it to 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; TOEFL CBT scores were highly but not perfectly correlated with GRE General Test scores for both groups of test takers; regressions of the General Test verbal scores on the TOEFL scores for ESL test takers were nonlinear, and the regressions of the other General Test scores were linear; trends in the variances of the General Test verbal scores associated with TOEFL scores were also nonlinear, and the trends were either unsystematic or negatively linear for the other General Test scores. 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

M. Usman Erdosy
Spring 2003

Variability in judgments of ESL compositions is inherent in the view that raters are “readers” with prior experiences. Such a view, however, obliges researchers to understand how personal

background and professional experience influence both scoring procedures and scoring criteria. These issues were explored by asking four raters to construct scoring criteria while assessing corpora of 60 TOEFL essays without the aid of a scoring rubric, and to discuss their procedures and criteria in follow-up interviews. 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, were undertaken to 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

Michael Rosenfeld, Philip K. Oltman, and Ken Sheppard
Spring 2003

The purpose of this study was to investigate the feasibility of two complementary approaches to assessing the validity of the TOEFL examination. One approach used evidence based on test content. In the context described in this report, evidence based on test content refers to the degree to which the items on the TOEFL examination are representative of the knowledge and skills required to demonstrate English proficiency in undergraduate and graduate programs throughout the United States and Canada. The content-oriented approach used in this pilot study involved item-rating procedures that were designed 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. The second approach used a criterion-related validation strategy. In this aspect of the study, 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.

New Monographs

MS – 24. Scales for Rating Language Performance: Descriptive Models, Formulation Styles, and Presentation Formats

Brian North
Spring 2003

This survey examines the description of language proficiency in assessment scales intended to guide the process of rating language performance. The aim is to provide developers of the new TOEFL with an overview of an underdocumented field. First the issue of scale purpose and descriptor content is discussed. Then the categories of description used in scales are classified and criticized. Ways of formulating distinctions between bands and prerequisites for a valid scale are explored. Finally, a comprehensive survey of different formats concludes with recommendations for the development of scoring rubrics for the new TOEFL speaking test.

MS - 25. Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus

Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay, and Alfredo Urzua
Spring 2003

To date, there have been few large-scale empirical investigations of academic registers, and virtually no such investigations of spoken academic registers. Given this lack of basic knowledge, it has been almost impossible to evaluate the representativeness of ESL/EFL materials and assessment instruments. Specifically in the context of the TOEFL 2000 effort, we have lacked the tools to determine whether the texts used on listening and reading exams accurately represent the linguistic characteristics of spoken and written academic registers.

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. Two major stages of analysis were completed: First, linguistic analyses of the text categories in the T2K-SWAL corpus were conducted to identify the salient patterns of language use in each academic register (across registers,

disciplines, and levels). Then, based on those findings, diagnostic tools were developed to indicate whether the language used in T2K Listening and Reading Comprehension tasks is representative of real-life language use.

Additional TOEFL 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 the performance of teaching assistants on the Test of Spoken English in relation to their classroom performance as judged by students; reports that the TSE test is a 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 the role of language proficiency as a moderator variable in assessing academic aptitude; a moderately strong correlation develops between

verbal aptitude tests in the 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 the 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 the 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 note taking 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 proficiency as 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. Prompt 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 that were 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 dialog 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; provides information on reader characteristics and their influence on ratings, 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 computer familiarity scale that was used to profile TOEFL examinees in terms of their computer familiarity (see TOEFL Research Report 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 Test of Spoken English. 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. A database of Test of Written English essays from several language groups was prepared and subjected to a number of computer analyses. Essays from English-speaking examinees were included to provide a baseline for comparison with the essays produced by examinees with English as a second language (ESL). Analyses revealed that topic differences influenced some of the essay variables, but the various language groups were not differentially affected. Language groups appeared to differ in the extent of directness, expressiveness, and academic stance of their writing styles. A number of computer-scored essay variables were related to official TWE scores assigned by human raters. In particular, number of words and the average length of words taken together are quite predictive of the 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. Now that TOEFL is computer-based, listening items are being created that include both audio and visual information. This study was conducted in order to begin to understand the effects of different types of visual presentations. The design examined 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 weakest of these interactions, type of stimuli by visual condition, was the most interesting and indicated that the presence of visuals results in facilitation of performance when the visuals bear information that complements the audio portion of the stimulus. The majority of the subjects indicated a strong preference for the presence of visuals.

RR – 67. Automatic Assessment of Vocabulary Usage Without Negative Evidence. Claudia Leacock and Martin Chodorow. November 2001. As part of the TOEFL program's effort to develop performance-based measures of communicative competence, we imple-

mented and evaluated an automated statistical method for assessing an examinee's use of vocabulary words in constructed responses. Our error-detection system, ALEK (*A*ssessing *L*exical *K*nowledge), 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. ALEK detects two types of errors: those that violate basic principles of English syntax (e.g., agreement errors as in **a desks*) and those that show a lack of information about a specific word (e.g., treating a mass noun as a count noun as in **a pollution*). The system evaluated word usage in essay-length responses to TOEFL prompts. ALEK was developed using 3 words and was evaluated on an additional 20 words that appeared frequently in TOEFL essays and in a university word list. The system performed with about 80% precision and 20% recall. False positives (correct usages that ALEK identified as errors) and misses (usage errors that were not recognized by ALEK) are analyzed, and methods for improving system performance are outlined.

RR – 68. Influence of Irrelevant Speech on Standardized Test Performance. Donald E. Powers, Wendy Albertson, Thomas Florek, John Malak, Kathy Johnson, Bill Nemceff, Mark Porzuc, Donna Silvester, Minhwei Wang, Richard Weston, Edward Winner, and Aleksander Zelazny. Winter 2002. The aim of this study was to (1) estimate the impact of distraction on test performance and (2) evaluate ways to reduce it. The distraction of interest was from fellow examinees taking a speaking test. Study participants were volunteers ($N = 171$) who had previously taken the Graduate Management Admission Test (GMAT), the Graduate Record Examinations (GRE) General Test, or the Test of English as a Foreign Language (TOEFL). They were invited to retake a different form of the same test under either distracting conditions or standard, distraction-free conditions. To reduce distraction, some participants used either headsets or headsets plus masking noise. Attempts to reduce distraction to an acceptable level were largely unsuccessful. The impact on actual test performance, however, was slight in the GMAT sample and negligible in both the GRE and TOEFL samples. The conclusion was that intermingling examinees with others who are taking a speaking test remains a concern, primarily because of strong negative perceptions by test takers. More effective means need to be devised to reduce or control distraction.

TOEFL 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 the clusters and test sections do not differ in their internal-consistency reliabilities and intercorrelations for the 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 item 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 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. A neural net approach was used 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. When various test sections were analyzed individually, the reliability of reading comprehension testlets was overestimated due to local dependence. Much less local dependence was found among listening comprehension testlets. It was concluded 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. Reading and writing, and listening and speaking items were combined 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. This paper 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. These ideas are illustrated in the contexts of the Graduate Record Examinations and language proficiency testing.

TR-15. Strengthening the Ties that Bind: Improving the Linking Network in Sparsely Connected Rating Designs. Myford and Wolfe. August 2000. The purpose of this study was to evaluate the effectiveness of a strategy for linking raters when large numbers of raters are involved in a scoring session and the overlap among raters is minimal. In sparsely connected rating designs, the number of examinees any given pair of raters has scored in common is very limited. Connections between raters may be weak and tentative at best. The linking strategy we employed involved having all raters in a Test of Spoken English scoring session rate a small set of six benchmark audiotapes, in addition to those examinee tapes that each rater scored as part of his or her normal workload. Using output from *Facets* analyses of the rating data, we looked at the effects of embedding blocks of ratings from various smaller sets of these benchmark tapes on key indicators of rating quality. We found that all of our benchmark sets were effective for establishing at least the minimal connectivity needed in the rating design to allow placement of all raters and all examinees on a single scale.

TR-16. Using a New Statistical Model for Testlets to Score TOEFL. Wainer and Wang. May 2001. Standard item response theory models fit to examination responses ignore the fact that sets of items (testlets) often are matched with a single common stimulus (e.g., a reading comprehension passage). In this setting, all items given to an examinee are unlikely to be conditionally independent (given examinee proficiency). Models that assume conditional independence will overestimate the precision with which examinee proficiency is measured. Overstatement of precision may lead to inaccurate inferences as well as prematurely ending an examination in which the stopping rule is based on the estimated standard error of examinee proficiency (e.g., an adaptive test). The standard three-parameter IRT model was modified to include an additional random effect for items nested within the same testlet. This parameter, γ , 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. The development and maintenance of item pools to support computer-based testing (CBT) programs have placed much greater demands on the item pretesting process than was the case with paper-and-pencil testing, and the TOEFL CBT is no exception. Of particular interest are procedures that might allow reduction in pretest sample sizes needed for IRT calibration purposes so that more items could be pretested given the fixed overall examinee volume that might be expected.

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-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-11. Technologies for Language Testing. Frase, Gong, 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. Computer-based speech technology, the capability of a computer system to accept and process spoken language, is considered a potentially super-enabling technology for computer users. Once a computer can adequately "understand" spoken language, the accessibility of computers increases by many orders of magnitude. As part of our ongoing effort to examine enabling and important technologies, we undertook with this study to 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. Our goal in this study was 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. In order to create credible forecasts for enrollment trends in intensive English language programs (IEPs) in the United States, analyses of past influences on IEP enrollments were undertaken. Reviews of available censuses of international students were carried out and related to the circumstances external to English programs that seemed to affect the movement of students into IEPs. The implications of similar circumstances in the future for IEP enrollments and for future IEP-related use of the TOEFL test were then posed. World economics, political developments, and educational policy and social change were viewed as the major categories of influences on IEP student flows. The major findings of the study suggest that IEP administrators should pay close attention to events occurring far outside the walls of the English program in order to anticipate future enrollments and to position their programs to respond to changes in the intensive English market. A final section discusses "test reverberation," or the attitudinal impact that changes in the TOEFL format will have on IEP students, instructors, and administrators.

MS-15. Washback in Language Testing. Bailey. June 1999. This monograph summarizes recent research on language testing washback. It begins by compiling several definitions of washback and related constructs. It then poses a model of language testing washback and examines the available research related to this model. The monograph concludes with recommendations for appropriate research methods to be used in future investigations of washback.

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 this first working framework is to guide the development of more specific frameworks and research agendas for the assessment of reading, writing, listening, and speaking, as both independent and integrated modalities.

The monograph is organized into five major parts. The first part presents a general introduction to the goals and key components of the project. The second part presents the historical background and work of the project leading to the development of this framework. The third and fourth sections present our conceptualization of a working framework that includes identifying the test domain, organizing the test domain, identifying task characteristics, identifying and operationalizing the variables, validating the variables, and building an interpretive scheme. The paper concludes with a discussion of the plans for proceeding with the work of the project.

MS-17. TOEFL 2000 Reading Framework: A Working Paper. Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, and Schedl. April 2000. The TOEFL 2000 framework paper (TOEFL Monograph 16) identifies a test domain and lays out a process for the design of a new TOEFL test based on communicative language abilities. 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. The reading framework described in this document was

developed by the authors — internal ETS staff and external reading experts — who have worked together over the past two years.

This monograph documents how three broad perspectives were considered in defining the construct of reading comprehension for assessment purposes: a processing perspective, a task perspective, and a reader purpose perspective. The reader-purpose perspective is recommended to guide the new test design for a number of reasons. One perceived advantage of this approach is that it is readily interpretable. It will be easier for test-score users, teachers, and examinees to understand how the construct is being defined. At the same time, the reader-purpose perspective is seen to be compatible with both the processing perspective and the task perspective.

Four purposes for reading in the academic context are identified: reading to find information, reading for basic comprehension, reading to learn, and reading to integrate information across multiple texts. These four reading purposes are seen to form a natural hierarchy that can serve as a basis for describing a continuum of reading proficiency. The first two purposes are addressed in the current TOEFL reading test format. The third and fourth purposes — reading to learn and reading to integrate information across multiple texts — would expand the construct being measured. Some tasks that might be used to assess reading for different purposes are described.

Finally, technological issues specific to the delivery of the reading test are described and a detailed research agenda related to the reading construct described in this document is provided.

MS-18. TOEFL 2000 Writing Framework: A Working Paper. Cumming, Kantor, Powers, Santos, and Taylor. April 2000. This paper builds on *TOEFL 2000 Framework: A Working Paper* (TOEFL Monograph 16) by setting out a preliminary working framework for the development of the writing assessment component of the TOEFL 2000 test.

The monograph is organized into four major parts. The first presents a conception of writing proficiency and focuses in particular on academic writing — the domain of the TOEFL 2000 test. The second part presents an initial writing framework for the test; it reviews the test domain, proposes an organizational scheme, and identifies the task variables of interest. The third section lays out an initial research agenda for establishing the validity of interpretations and the appropriateness of action that would result from the introduction of writing measures growing out of this framework and approach to test design. The paper concludes with a discussion of the important ways in which the TOEFL 2000 approach to testing writing is intended to improve on its predecessors.

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 *TOEFL 2000 Framework: A Working Paper* (TOEFL Monograph 16).

This monograph consists of six sections. After a brief introduction, an overview of academic listening is presented that outlines theory and research in the areas of listening in general and academic listening in particular. The third section describes theory and research on the variables that characterize academic listening and that might drive the difficulty of a listening task.

The fourth section addresses technological issues for the listening section of the TOEFL 2000 test. A research agenda is presented in the fifth section of the monograph, and the final section describes the features that distinguish the TOEFL 2000 test from its predecessors.

MS-20. TOEFL 2000 Speaking Framework: A Working Paper. Butler, Eignor, Jones, McNamara, and Suomi. June 2000. This paper was prepared based on an initial overall framework paper developed for the TOEFL 2000 project by Jamieson, Jones, Kirsch, Mosenthal, and Taylor (TOEFL Monograph 16). The paper applies concepts advanced in the overall paper to the modality of speaking. In doing so, this document presents an initial framework for research and development activities for the speaking component of the TOEFL 2000 test. The paper should be viewed as a work in progress, as research activities presently under way for TOEFL 2000 will undoubtedly bring about refinements to the contents of this document.

The paper is made up of six parts. Part 1 provides an introduction to the overall document. In Part 2, oral discourse is discussed from a sociological perspective as well as in terms of speech act theory. Part 3 discusses the details of the speaking framework for the TOEFL 2000 test. Identification of the test domain and relevant task characteristics and variables is discussed, along with some of the factors suspected to influence the difficulty of speaking tasks. Part 4, some of the technical issues involved in eliciting and capturing speech samples are discussed. Part 5 contains a list of relevant research activities that should be pursued as the project progresses. The final part 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. Michael Rosenfeld, Susan Leung, and Philip K. Oltman. November 2001. The primary purposes of this project were to (1) aid in translating or operationalizing the theoretical frameworks that were developed in reading, writing, speaking, and listening by the TOEFL framework teams into task statements that undergraduate and graduate students would need to perform in order to complete their academic programs; (2) have these statements reviewed and evaluated by undergraduate and graduate faculty experienced in teaching nonnative speakers of English, as well as by undergraduate and graduate students who were nonnative speakers of English through the use of survey instruments; (3) provide analyses of these results to aid in the design of test specifications and assessment measures for the new TOEFL; and (4) document these results to assist in supporting the validity of the new TOEFL.

MS-22. Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation Into Raters' Decision Making, and Development of a Preliminary Analytic Framework. Alister Cumming, Robert Kantor, and Donald E. 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. Phase 1 developed empirically (a) an initial framework to describe the decision-making behaviors of 10 experienced ESL/EFL instructors/assessors of differing backgrounds who produced concurrent verbal reports while each rating 60 TOEFL essays, and (b) a questionnaire to profile relevant variables in the raters' backgrounds. Phase 2 refined the framework, gathering additional data from seven highly experienced English-mother-tongue composition assessors while they rated 40 TOEFL essays and from seven of the same experienced ESL/EFL instructors/assessors while they rated 30 ESL compositions for five prototype tasks that involve writing in response to reading or listening material.

MS-23. The Effects of Notetaking, Lecture Length, and Topic on the Listening Component of TOEFL 2000. Patricia L. Carrell, Patricia A. Dunkel, and Pamela Mollaun. Winter 2002. The present study examined the effects of notetaking, lecture length, and topic, as well as two aptitude variables, on listening comprehension with ESL students representative of the TOEFL population. A total of 234 ESL students at five participating universities in the United States took a computer-based listening comprehension test, a short-term memory test, the listening comprehension section of a disclosed Institutional TOEFL, a debriefing questionnaire, and a biodata questionnaire. 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.

