ETS TOEFL

Monograph Series

MS - 26
February 2005

A Teacher-Verification Study of Speaking and Writing Prototype Tasks for a New TOEFL

Alister Cumming
Leslie Grant
Patricia Mulcahy-Ernt
Donald E. Powers

A Teacher-Verification Study of Speaking and Writing Prototype Tasks for a New TOEFL

Alister Cumming
University of Toronto, Ontario, Canada

Leslie Grant
University of Colorado at Colorado Springs

Patricia Mulcahy-Ernt
University of Bridgeport, CT

Donald E. Powers
ETS, Princeton, NJ

RM-04-05

ETS is an Equal Opportunity/Affirmative Action Employer.

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

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

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Foreword

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

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

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

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

Monographs 16 through 20 were the working papers that laid out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, described the process by which the project was to move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extended the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). These conceptual frameworks guided the research and prototyping studies described in subsequent monographs that resulted in the final test model. The culmination of the TOEFL 2000 project is the next generation TOEFL test that will be released in September 2005.

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

TOEFL Program
Educational Testing Service

Abstract

This study was undertaken, in conjunction with other studies field-testing prototype tasks for a new Test of English as a Foreign Language™ (TOEFL®), to evaluate the content validity, perceived authenticity, and educational appropriateness of these prototype tasks. We interviewed 7 highly experienced instructors of English as a second language (ESL) at 3 universities, asking them to rate their students' abilities in English and to review samples of their students' performance to determine whether they thought 7 prototype speaking and writing tasks being field-tested for a new version of the TOEFL test (a) represented the domain of academic English required for studies at English-medium universities or colleges in North America, (b) elicited performance from their adult ESL students that corresponded to their usual performance in ESL classes and course assignments, and (c) realized the evidence claims on which the tasks had been designed. The instructors thought that most of their students' performances on the prototype test tasks were equivalent to or better than their usual performances in classes. The instructors viewed positively the new prototype tasks that required students to write or to speak in reference to reading or listening source texts, but they observed certain problems with these novel tasks and suggested ways that the content and presentation might be improved for the formative development of these tasks.

Key words: TOEFL, test validity, test content, authenticity, instructors' perceptions, integrated language tasks

Acknowledgments

We thank the participating instructors for their integral contributions to this research, Mary Enright for helpful advice at several stages of the study, Antony Kunnan for useful comments when this report was presented in a colloquium at the Annual TESOL Convention in Salt Lake City (April 11, 2002), Altamese Jackenthal and other ETS staff for preparation of the stimulus materials and transcriptions, as well as graduate students Keanre Eouanzoui, Joshua Hogland, Elizabeth Lyons, Linda Steinman, and Teya Dodge Yu for their assistance in the transcription of interviews, compilation of data, or statistical analyses.

Table of Contents

List of Tables

Page

Introduction

The present study was one of many coordinated studies aimed at developing new task types for the Test of English as a Foreign Language™ (TOEFL®) following the conceptual foundations for new integrated reading, writing, listening, and speaking tasks with communicative, academic orientations for the test outlined in Jamieson, Jones, Kirsch, Mosenthal, and Taylor (2000) and elaborated in more detail for each mode of communication in Bejar, Douglas, Jamieson, Nissan, and Turner (2000, for listening); Butler, Eignor, Jones, McNamara, and Suomi (2000, for speaking); Cumming, Kantor, Powers, Santos, and Taylor (2000, for writing); and Enright et al. (2000, for reading). Specifically, the present study was coordinated with field tests of prototype tasks for the new TOEFL in the autumn of 2000 that involved 475 ESL students in Australia, Canada, Hong Kong, Taiwan, and the United States (described in Enright & Cline, 2002). The present study supplemented this larger field test by gathering in-depth data from a small sample of highly experienced English as a second language (ESL) instructors, whom we interviewed about their perceptions of the prototype tasks and of their students' performance on these tasks in the field test. The present research complemented other studies of student performance on the field test, such as analyses of the difficulty, score reliability, and consistency of measurements of the tasks (Enright & Cline, 2002) or analyses of the relationship of scores on the field tests to external criteria, including placement in ESL classes, scores on other ESL tests, teachers' ratings of students' English proficiency, and students' self-ratings of their English proficiency (Bridgeman, Cline, & Powers, 2002).

The purpose of the present study was primarily formative. The research was undertaken to inform decisions about the content validity, authenticity, and educational appropriateness of the prototype tasks under consideration for a new version of the TOEFL test. In addition to verifying these qualities of the prototype tasks, we aimed to produce findings that could help to guide revisions to certain task types or decisions about deleting some of them from the pool of task types under consideration for the new test. We looked to experienced ESL instructors to provide information that would help to answer the following research questions:

1. Is the content of the prototype tasks being field tested for the new TOEFL perceived to assess the domain of academic English required for studies at an English-medium university in North America?

2. Is the performance of ESL students on the prototype tasks field-tested for the new TOEFL perceived to correspond authentically to the performance of these students in their ESL classes?

3. Are the prototype tasks being field-tested for the new TOEFL perceived to realize the evidence claims on which the tasks were designed?

The research questions address the concepts of content validity, authenticity, and educational relevance. These concepts have been principal foci of criticisms about the task types in the current TOEFL test. They have therefore formed a major impetus for revisions to the test that are now underway. They have also formed the basis for new definitions of the constructs to guide the design of a new version of the test (Jamieson et al., 2000). For example, Chapelle, Grabe, and Berns (1997) argued that the content and format of the TOEFL needed to be revised substantively to reflect contemporary concepts of communicative competence and practices of language teaching and learning. Likewise, Hamp-Lyons and Kroll (1997) and Raimes (1990) criticized the current writing component of the TOEFL (or the Test of Written English™) for not assessing the types of writing that students realistically need to perform at universities in North America and for the test not having educationally relevant definitions of the construct of writing ability in a second language. Because the current TOEFL does not engage examinees in actually speaking English in any way, various arguments have been made to add a component to the test that assesses speaking abilities directly (Butler et al., 2000). Perhaps the strongest criticism of the task types in the current TOEFL is that many items focus on discrete knowledge about the forms of the English language (such as multiple-choice items about grammar or vocabulary) that can easily be coached or can have a negative washback on learning and teaching English by directing students' attention to learning such items rather than developing their abilities to communicate proficiently (Alderson & Hamp-Lyons, 1996; Bailey, 1999).

The design of prototype tasks for the new TOEFL test has tried to address these concerns by creating a range of new task types for the assessment of writing and speaking, by defining precisely the constructs that these tasks are intended to assess, and by integrating the modalities for language production (i.e., writing and speaking) with tasks that involve examinees reading or listening to source texts, as they would in real academic tasks in English.

The rationale for the present study also follows from recent, expanded conceptualizations of the centrality of construct validity in test development, requiring diverse kinds of evidence from multiple relevant sources about the interpretation of test scores and the uses to which test results are put (e.g., as argued by Messick, 1989; Moss, 1992). This view has proved to be as important for language testing as for other domains of assessment (Bachman, 2000; Cumming, 1996; Fulcher, 1999; Kunnan, 1998). For instance, many educators have argued that the content of tests in education should relate integrally to curriculum standards as well as to students' typical achievements and performance requirements. This perspective has promoted the use of task types for assessments that authentically resemble the tasks that students actually need to perform in their academic studies (Darling-Hammond, 1994; Freedman, 1991; Gipps, 1994; Linn, Baker, & Dunbar, 1991). For tests of language abilities, many theorists have recently argued that assessment tasks need to "somehow capture or recreate…the essence of language use" (Bachman, 1990, p. 300), urging that authenticity should be a fundamental criterion for the validity of tasks that aim to evaluate how well people can really communicate in a language (Bachman & Palmer, 1996; Lewkowicz, 2000; Spence-Brown, 2001; Spolsky, 1985). In turn, to achieve educational relevance, test tasks should elicit performance that is congruent with instructors' perceptions of their students' abilities, as numerous different inquiries in second-language assessment have recently emphasized and demonstrated (e.g., Brindley, 1998, 2000; Chalhoub-Deville, 1995; Elder, 1993; Epp & Stawychny, 2001; Grant, 1997; North, 1995, 2000; Stansfield & Kenyon, 1996; Sullivan, Weir, & Saville, 2002; Teachers of English to Speakers of Other Languages, 1998). Nonetheless, surprisingly few studies (apart from those cited in the preceding sentence) have systematically sought the input of experienced teachers for the design or validation of second-language tests, despite a published literature that has established the reliability and informative value of teachers' judgments of their students' abilities or achievements in relation to standardized tests in various domains of education (e.g., Griffin, 1995; Hoge & Coladarci, 1989; Miesels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Sharpley & Edgar, 1986).

With these ideas in mind, we initiated this study to explore how a small group of highly experienced ESL instructors might evaluate the content of the prototype tasks being field-tested for the new TOEFL, judge the authenticity of their students' performance on these prototype tasks, and evaluate whether the prototype tasks appropriately fulfilled the purposes for which

they had been designed. We solicited the participation in three sites (Central Michigan University, University of Bridgeport, and University of Toronto) of experienced ESL instructors who had taught ESL students who had participated in the previous months in the field-testing of prototype tasks for the new TOEFL. We met with these instructors in December 2000 and January 2001 to review and evaluate samples of their students' written and oral performance on prototype tasks during the field test of these tasks in the autumn of 2000. We also asked the instructors to complete questionnaires profiling their teaching experiences and to evaluate the effectiveness of the various task types in the prototype test. Finally, we asked the instructors to rate and describe the abilities of a sample of their current ESL students, then to provide examples of the students' work from their ESL classes.

Method

Participants

The participants in this research were four instructors at the University of Toronto, two instructors at Central Michigan University, and one instructor at the University of Bridgeport. We solicited all seven through referrals from local ESL program administrators, aiming for a purposeful sample of instructors who were highly experienced in ESL/EFL instruction, had relevant educational qualifications, and had devoted their careers to this field. We focused on these three universities knowing that they had ESL programs (and a range of experienced instructors working in them) from which a sufficiently large number of ESL students (either destined for, or already participating in, academic courses at English-medium North American universities) had volunteered to participate in the field tests of new TOEFL prototype tasks in the autumn of 2000 to make the inquiry feasible (i.e., for instructors to be able to consider their current students' performances in the field tests). At each of these three universities we circulated letters of invitation to all ESL instructors whose students had participated in the field test of the new TOEFL tasks, offering them a stipend of $100 to meet with us for the interviews and to complete the questionnaires. We accepted all instructors who volunteered to participate, which turned out to be about half of the full-time, experienced ESL instructors teaching university ESL credit or preuniversity ESL courses at each institution. Although we achieved some geographic distribution in this selection of sites, we were not able to span the geographic spread of the larger field-testing study (i.e., participants outside North America were not

solicited), nor could we claim that the instructors or their students were in any way representative of the population of people who usually take the TOEFL or who study or teach ESL either in North America or internationally. Each instructor completed a profile questionnaire (displayed in the appendix). Their individual characteristics are presented in aggregate here, however, to preserve their confidentiality.[1]

The seven instructors had on average 11 years of experience teaching ESL, ranging from 8 to 17 years per individual. Each instructor had made English language teaching his or her principal career, and five had also worked as assessors of English at their employing institutions. Six instructors had master's degrees, and one instructor had a doctorate in foreign language education; in addition, five instructors had certificates in teaching English as a second language. Each instructor's dominant and principal language was English, though two instructors had been raised in families that spoke another language as well, and one of them presently used another language at home. They ranged in age from 33 to 60, were four females and three males, and considered themselves as either "competent" (four people), "expert" (one person), or in between "competent" and "expert" (two people) with respect to their skills in assessing English language proficiency. At the time of the interviews, each instructor had been teaching his or her respective classes for either 14, 15, 16, or (in the case of Bob, described below) 35 weeks, and they considered that they knew their individual students' abilities in English well, as indicated by self-ratings of mostly 4 or 5 on a scale of 5, where 5 = very well and 1 = not well. However, four instructors rated their perceived knowledge of students' reading, listening, and speaking abilities as being lower than their knowledge of students' writing abilities. Two instructors rated their knowledge of students' speaking or reading as 2 or 3 (on the scale of 5), reflecting the types of regular interactions they had with these students and the skills they emphasized in the courses they taught. For example, an instructor who only taught writing to these students twice per week responded with a 5 when rating knowledge of the students' writing abilities but only a 2 when rating knowledge of students' listening abilities. Nonetheless, three instructors rated their overall knowledge of their students' English abilities at 5.

Teaching Contexts

At the University of Toronto, two instructors, who gave themselves the pseudonyms Sara and Tracy, were teaching different sections of an intensive (i.e., daily classes of several hours' duration), 15-week, noncredit ESL course for students preparing to enter the university in the

following year. The students' TOEFL scores were around 500 on the pencil-and-paper or institutional version of the test (the university's minimum level for admission was 550 to 600, depending on the program). Sara's students were grouped into a class with a level of English proficiency just below that of the students in Tracy's class. The course focused on academically oriented reading, writing, speaking, and listening tasks in preparation for university studies. Two other instructors, "Francis" and "Tyler," team-taught a course for ESL students already admitted to programs in various departments in the university but who wanted to improve their English abilities. One instructor focused on oral abilities, while the other focused on literate abilities, but both instructors taught the same students. The course was held 3 hours per week over the academic term of 14 weeks; each instructor was responsible for 1.5 hours of instruction weekly. The students received credit on their academic transcripts for their participation in the course, but the course did not carry academic credit toward any degree programs. All four of these ESL classes were extremely diverse in respect to students' ethnic and linguistic backgrounds and their fields of academic study.

The two instructors at Central Michigan University, "Steve" and "Ann," taught in the Dual Program in the English Language Institute, a program for students who have TOEFL scores between 500 and 550 and do not meet regular university admissions requirements (which is most often TOEFL scores of 550 or higher). These students usually take courses both in English and in their field of study (thus, the name *dual*); the English courses, however, are not credit bearing. Individual course assignments varied among these dual program teachers, but each course met twice weekly for classes of 1.5 hours. Steve and Ann both taught writing courses that emphasized the writing of academic papers, particularly through using reference sources. But Ann's students had a higher level of English proficiency than Steve's did; often students took Steve's course first and then opted for more writing in Ann's course. Ann also taught a course in the dual program that focused on listening to academic lectures and taking appropriate notes.

At the University of Bridgeport, "Bob" taught graduate-level courses for international students in a year-long program for teacher candidates seeking a master of science in education and academic preparation for teaching English in their native countries. All their coursework was completed in English over a three-semester program, which began during the summer prior to our data collection in the autumn of 2000. Screening for admission into the program included passing scores on the university's English language assessment battery or a minimum score of

550 on the TOEFL. Bob was familiar with the full range of his students' performance in English because by the time of the interviews he had observed the students during several of the graduate classes he taught, during their field placements, in the context of independent writing projects, and during university community activities. Each student completed approximately 5 hours of instructional classroom and tutorial time per week with Bob. The faculty-to-student ratio of 1 to 12 in this program gave Bob the opportunity to observe these students in a variety of academic and conversational speaking and listening situations and to see the academic writing and reading performances of each student.

Interviews

We interviewed instructors shortly after their ESL courses were completed in either December 2000 or early January 2001, once data on their students' performance on the prototype tasks (from the autumn of 2000) had been processed and made available to us in CD-ROM form. All instructors were interviewed individually except for Francis and Tyler, who met for their interview together because they had taught the same students and had complementary perspectives on either the oral or the literate abilities of the same ESL students. Each interview lasted from 2.5 to 4 hours, was audiotaped, and was later transcribed in full. Bob returned for a second interview to complete the descriptions of the performance of the 10 students he had taught.

Each interview followed a common format. Instructors were asked to describe the ESL course they had been teaching, the sample of students' work they had brought to the interview, as well as the ratings they had given previously (as part of the larger prototype study in the autumn 2000; see Enright & Cline, 2002) to the writing and speaking abilities of the focal students they had selected from their classes. Then, for each of the focal students, the instructor either read or listened to the student's writing or speaking performance on the prototype TOEFL tasks, commenting on how representative they thought the performance was for the student, based on their knowledge of that student and the person's typical performance in class and in course assignments. Each student's performance was provided on a CD-ROM from ETS; when students had written by hand, photocopies of their writing were given to the instructors to review. In the final part of the interview, the instructors individually completed the profile questionnaire shown in the appendix.

Questionnaire

We developed, piloted, and then refined a questionnaire (shown in the appendix to this report), which asked each instructor to evaluate the effectiveness (on a scale from 1 = not effective to 5 = very effective) of each of the prototype task types; then to explain briefly why the instructors considered each task type effective or not; then to rate how well they thought the tasks "represent the domain of academic English required for studies at an English-medium university" and to explain why which tasks do this best, and what might be missing from the tasks. The final section of the questionnaire asked the instructors to profile themselves, their teaching and assessment experiences, and their knowledge of the ESL students they had been teaching.

Samples of Students' Performances

To keep each interview to a reasonable duration (i.e., no more than 2 hours per instructor), we decided in advance that we would have to sample from among a total of the 13 integrated tasks (i.e., tasks involving more than a single language modality) and 6 independent tasks (i.e., those that involved a single modality) that were administered in the field test, focusing particularly on tasks that we collectively had decided might best display students' abilities. The seven tasks (and the constructs they were intended to evaluate in students' performance, that is, the evidence claims for the tasks) we considered were

- independent writing (TOEFL essay), intended to assess writing about personal experiences and opinions;
- academic writing in response to a lecture (on plate tectonics), intended to assess writing about academic topics;
- academic writing in response to a reading passage (on dance), intended to assess writing about academic topics;
- independent speaking, intended to assess speaking about personal experiences and opinions;
- speaking in response to a lecture (on ground water), intended to assess speaking about academic topics;

- speaking in response to a conversation (about student housing), intended to assess speaking in academic contexts; and
- speaking in response to a reading passage (about innate vs. learned abilities), intended to assess speaking about academic topics.

This selection of tasks from among the many prototype tasks that were field-tested provided a range of one of each of the three different types of writing and the four different speaking types of tasks. But the sample of tasks was not sufficient to permit our analyses to be able to distinguish between the task type, the stimuli material, or the topic. We have not included descriptions of these tasks here for two reasons: (a) test security and (b) most of the tasks are being revised substantively, on the basis of analyses from the present and other field tests, so we would not want, by publishing them here, to mislead future users of the test about what the content or format of the tasks will be in the new TOEFL. Parallel to these prototype tasks, the instructors also provided ratings (as part of the larger prototype study), using a scale from 1 ("is clearly insufficient") to 5 ("is highly developed, will not be a barrier to success"), of each of the focal students' abilities in "writing about personal experiences and opinions," "writing about academic topics," "speaking about personal experiences and opinions," and "speaking about academic topics" (i.e., corresponding to the evidence claims for these tasks). We only considered the students' performance on writing and speaking tasks (most of which included tasks that involved reading and listening stimuli), because these provided holistic samples of students' performance in extended discourse. We did not think that the instructors could readily judge students' performance on closed-format item types, such as multiple-choice or other formats, that typically appeared in the assessments of their reading and listening abilities. The instructors also rated the students' listening and reading abilities, but we did not interview the instructors about these abilities (for reasons cited above). See Bridgeman et al. (2002) for analyses of these data for the full set of participants in the field test.

Summary of Data

In sum, the data we gathered data were from seven ESL instructors at three universities in the United States and Canada, who

- in the autumn of 2000 had rated the English proficiency of their students who had participated in the field test of prototype tasks for the new TOEFL, and then
- met with us in December 2000 and January 2001
    - to complete questionnaires profiling their professional qualifications,
    - to evaluate prototype writing and speaking tasks for the TOEFL generally and in respect to the evidence claims on which the tasks had been designed, and
    - to rate the performances on three prototype writing tasks and four prototype speaking tasks of between 4 and 10 of the ESL students whom each ESL instructor had taught over the previous academic term, and who had participated in the field test of the prototype TOEFL tasks in the autumn of 2000.

We also had test scores on the prototype tasks in the field test for the students whose performance the instructors had evaluated in the autumn of 2000 during the field test.

Analyses

To answer our research questions, we developed a common framework for tallying and reviewing the data from the questionnaire and transcribed interviews. We then conducted initial analyses individually at each site (i.e., by the interviewer who had conducted the original interview), preparing three separate reports for each of the three sites. Then we shared among ourselves the full set of data and our respective interpretations, synthesizing these interpretations into the present, common report. For items where the instructors had made numerical evaluations, we tallied these. But most of the data involved open-ended responses either to questionnaire items or to the questions posed in the standard interview schedule. We interpreted these using a constant-comparative method (Miles & Huberman, 1994), reviewing the data to identify trends and themes and then crosschecking them (against the data and with each other's interpretations of them).

To answer our first research question (Is the content of the prototype tasks being field-tested for the new TOEFL perceived to assess the domain of academic English required for studies at an English-medium university in North America?), we analyzed the instructors' responses to Item I.3 in the questionnaire (see the appendix). To answer our second research question (Is the performance of

ESL students on the prototype tasks field-tested for the new TOEFL perceived to correspond authentically to the performance of these students in their ESL classes?), we categorized and analyzed the instructors' spoken judgments of whether they thought their students' performance on the various task types corresponded to their usual performance in classes, identifying trends in their stated impressions about how well or poorly their students had performed.

For the third research question (Are the prototype tasks being field-tested for the new TOEFL perceived to realize the evidence claims on which the tasks were designed?), we analyzed the instructors' responses to Items 1 and 2 (and their corresponding subsections) on the questionnaire, distinguishing between the instructors' evaluations of the various kinds of writing and speaking tasks. We also calculated Spearman's rank-order correlations (*rho*) between (a) the ratings by the seven instructors of their students' speaking and writing abilities in English (i.e., according to the evidence claims cited above) and (b) the scores these students received in the field test (ETS, 2001) on the seven prototype tasks that intended to measure the constructs corresponding to the abilities that the instructors had rated. These items on the questionnaire and the separate ratings by teachers followed the definitions of the constructs for reading and writing tasks specified in Rosenfeld, Leung, and Oltman (2001) as well as in the frameworks prepared by Butler et al. (2000, for speaking) and Cumming et al. (2000, for writing). For these analyses, in addition to *p* values, we estimated effect sizes using the Aaron-Kromrey-Ferron formula,[2] appropriate to this small, disparate sample. As described above, data were missing for numerous task performances for certain students, so the number of cases varied for each correlation; we omitted the missing data from the correlation analyses.

Results

After first describing the sample of students' performances we obtained, we report results below in respect to the three research questions guiding the study. During the interviews, the instructors considered the task performance of samples of 4 to 10 of their students, whom they had selected to represent the range of students' abilities, gender, and cultural backgrounds in their classes and whose English abilities they were confident they were familiar with. The available sample of data from ESL students was restricted because only a portion of each instructor's students had originally volunteered to participate in the field test, because we wanted to keep the interviews with each instructor within reasonable durations of time (i.e., no more than

2 hours), and because there were large amounts of missing data that resulted from mechanical problems during the field test. Over half the spoken data were missing because of technical difficulties in transmitting these during the field-testing (i.e., for almost all students in Bridgeport, for about half the spoken responses in Toronto, and for a few students in Central Michigan), and some of the audio recordings were of such poor quality that the instructors could not evaluate them. Some students wrote on computers, whereas others wrote their responses by hand. A small number of the writing tasks were not available for the interviews, either for technical reasons or because the students did not complete particular tasks. Table 1 shows the numbers of students and the distribution of each task type that each instructor commented on as well as the proportion of missing data (i.e., about one fifth of the overall set of student performances on the field test). Our rationale for continuing with the study, despite these large losses of data, was that we did have samples of data across the full range of task types for each instructor's students as well as the opportunity to obtain in-depth data from each instructor even though we lacked the breadth of data we had originally sought.

Table 1

Number of Student Performances on Three Writing and Four Speaking Tasks Commented on by the Seven Instructors and Percentages of Missing Data for Each Task Type and Instructor

Instructor	Students considered	Writing tasks[a]	Speaking tasks[b]	Missing data
Tracey	6	18	5	45%
Sara	4	12	5	39%
Tyler	6	15	11	38%
Francis	6	15	10	40%
Ann	6	18	23	2%
Steve	6	18	24	0%
Bob	10	30	4	51%
		Totals		
7	44	126	82	22%

[a] 4% missing. [b] 54% missing.

Research Question 1: Content Validity

In considering the prototype tasks overall, all of the instructors expressed in their questionnaire responses (after hearing samples of their students' speech and reading samples of their writing on the seven tasks) that they thought these tasks represented quite well (i.e., mostly ratings of 4 out of 5) the domain of academic English required for studies at an English-medium university. The integrated nature and realistic academic content of the reading, listening, speaking, and writing tasks were the chief reasons the instructors cited. For example, one instructor observed, "If a student can read/listen to a fairly complicated text and respond clearly, it seems like a good indication they'll be able to do academic work." Another instructor noted, "They all seem to be typical academic tasks, as realistic as can be in a testing situation." Another wrote in her overall evaluation, "Both the simulated lectures and the readings were close to real university-level material." Likewise, another instructor judged that the tasks "cover skills needed by students, i.e., attending to lectures, writing essays, exams, interpreting lectures/text books." One instructor, however, rated the tasks a bit lower, giving an overall rating of 3 out of 5. She noted that the tasks might be appropriate for undergraduate students, but she was concerned about graduate students: "However, for graduate students, they would not necessarily have the language ability needed to deal with these topics, but might have the language ability needed to succeed in their particular fields." The instructor who taught solely at the graduate level, though, rated the tasks overall as 5 but indicated he would like to see more open-ended choices "based on individuals' preferences, interests, and inclinations."

When asked which tasks best represent the domain of academic English required for university studies, five of the seven instructors indicated the written and spoken responses to the readings and lectures. One instructor said she found the spoken responses were "less representative" than written responses of university studies; and one liked the *idea* of spoken responses to the lecture, but felt the content of the lectures (and readings) was difficult.

In addition to these positive comments, the instructors expressed some reservations about the content of the prototype tasks, particularly when asked what they thought might be "missing" from the tasks. Several instructors suggested they thought the tasks could be more cognitively or intellectually challenging or complex. For instance, one instructor thought the tasks "reflect stressful (e.g., exam) situations, but don't allow for the application of some study skills and strategies which take time or relaxation." In a similar vein, another instructor observed that the

tasks "don't include anything requiring the students to apply or extend—i.e., think about, around, or beyond—the material." A third instructor similarly suggested that "some tasks could ask a student to expand on a topic and/or provide analogies/contrasts to the theories or information presented." Another cautioned that the tasks "don't necessarily give us a clear idea of their speaking/writing ability." The brief time allocated to do the tasks was a concern for several of the instructors, though they acknowledged the difficulties of overcoming this in the context of a standardized test. One teacher noted (as described above) the difficulty that graduate students might have with "general interest" topics, such as geology, dance, and ecology, and suggested "incorporating a wider variety of discipline-specific topics (e.g., business topics)." Another instructor likewise thought there should be more choices for the examinees to write about based on their individual preferences or interests. One instructor also observed that the processes of conversation in classroom situations, such as "interacting, interrupting, holding the floor," were not assessed. Two instructors pointed out that visual material, such as graphs, pictures, and outlines, were limited in the lectures and readings. One comment suggested that if the prompts for the reading passages were presented both prior to and after the reading tasks, this would approximate conditions for reading in academic contexts.

Research Question 2: Authenticity of Performance on the Prototype Tasks

The instructors verified during their interviews that most of their students' performances on the prototype TOEFL tasks corresponded closely to how these students usually performed on classroom tasks or course assignments in their ESL courses. For the 208 performances of their students on the prototype tasks that the instructors evaluated, they stated 70% of the time that students' performances were equivalent to how the students usually spoke or wrote English in their ESL classes. Moreover, for an additional 8% of the ratings the instructors thought the students did better than they usually did in their ESL classes. But for 22% of the ratings the instructors thought the students did worse than they usually did in classes. These percentages were relatively consistent across the seven instructors and the three sites, as shown in Table 2 and listed in detail for each type of writing task in Table 3 and each type of speaking task in Table 4. Most of the estimates of the students performing worse on the prototype types involved three tasks, suggesting that these tasks might usefully be revised for or deleted from future field tests: the reading-writing task on dance, the listening-writing task on plate tectonics, and the

reading-speaking task on innate versus learned behavior. The listening-writing task on plate tectonics was seemingly the most problematic of these tasks, as 46% of the students were perceived to perform worse on this task than they usually did in their classes.

Table 2

Instructors' Evaluations of Correspondences Between Students' Performance on Seven Prototype Tasks and Their Usual Performance in Classes

	Toronto	Michigan	Bridgeport	Total for 3 sites
Better than class performance	9%	7%	9%	8%
Matches class performance	71%	70%	67%	70%
Worse than class performance	20%	23%	24%	22%

Table 3

Numbers of Students' Performances on Each Writing Task Evaluated by the Seven Instructors at Three Sites

	Toronto	Michigan	Bridgeport	Total for 3 sites
Independent writing tasks judged in total	16	12	10	38
Better than class performance	3	2	1	6
Matches class performance	12	10	9	31
Worse than class performance	1	0	0	1

(Table continues)

Table 3 (continued)

	Toronto	Michigan	Bridgeport	Total for 3 sites
Listening-writing (plate tectonics) tasks judged in total	13	12	10	35
Better than class performance	0	0	0	0
Matches class performance	8	6	5	19
Worse than class performance	5	6	5	16
Reading-writing (dance) tasks judged in total	17	12	10	39
Better than class performance	3	0	1	4
Matches class performance	10	7	7	24
Worse than class performance	4	5	2	11

Table 4

Numbers of Students' Performances on Each Speaking Task Evaluated by the Seven Instructors at Three Sites

	Toronto	Michigan	Bridgeport	Total for 3 sites
Independent speaking tasks judged in total	13	12	1	26
Better than class performance	0	1	0	1
Matches class performance	11	8	0	19

(Table continues)

Table 4 (continued)

	Toronto	Michigan	Bridgeport	Total for 3 sites
Worse than class performance	2	3	1	6
Listening-speaking (groundwater) tasks judged in total	10	12	1	23
Better than class performance	1	0	0	1
Matches class performance	9	11	1	21
Worse than class performance	0	1	0	1
Listening-speaking (housing) tasks judged in total	9	11	1	21
Better than class performance	1	2	1	4
Matches class performance	8	9	0	17
Worse than class performance	0	0	0	0
Reading-speaking (innate vs. learned behavior) tasks judged in total	13	12	1	26
Better than class performance	0	1	0	1
Matches class performance	7	7	1	15
Worse than class performance	6	4	0	10

The interview transcripts were full of expressions that attest to the instructors' perceptions that the prototype tasks tended to elicit performances from their students that they readily recognized as typical of these individual learners' performances in English. For instance, in the independent speaking or writing tasks:

> That's our student. That's typical of him at his best. He can be very articulate and quite forceful, and a strong speaker, and he knows what he wants to say, quite often. That's more like what we usually saw, as I say, when he was on his best form. (Francis, student 130075, Independent speaking task)

> Fairly close. Her pronunciation is the same, she was a little bit hesitant in class. She sounds a little bit nervous, and that corresponds with, like, when she had to talk in front of the whole class. She had that same kind of nervousness in her voice. When she was involved in conversations with groups, she was a little bit more fluent, but that seems to correspond. (Sara, student 130016, Independent speaking task)

> Her first paragraph is indicative of the type of work she does in class. She is a free thinker, and she answers that well. And in her second paragraph she gives a good reason for connecting her activities with the global marketplace, I guess. And the last paragraph is more like a conclusion. Yeah—This is good; this is good. I like the way she writes here. This is typical of her in the class. (Bob, student 240081, Independent writing task)

Such remarks also abounded as the instructors commented on the match between a student's performance on the prototype tasks that integrate language modalities and the instructor's knowledge of the student's language abilities. For example:

> I think that sample really demonstrated her ability well. Uh, she answered the prompt so well and identified all the parts and gave a lot of details from the lecture, and I thought it was really good the way the she connected the ideas together using transitions. Uh, so she was not just repeating facts, but making a synthesis of the ideas. So I think that sample was really good; it represented her well. (Ann, student 240022, Groundwater listening-speaking task)

> Ok, I think she did a very good job with this response. The main idea in the reading is the difference between the influence of men and the control of men and the influence of women, and I think she talks about that quite a bit, so I think I am pretty impressed with her response to this one. It may be something that she is more familiar with than plate tectonics, uh, but I think she did a good job, and I think this is the kind of writing that she is capable of doing. (Steve, student 240060, Dance reading-writing task)

> That is typical of him. He tries to come up with a common denominator. Even though directions in class are given to him, and he has to re-do, at times, he is under the influence of the way he thinks rather than the way he gets directions. Yes, that is representative of what he does in class. The difference, of course, is in class, when he does respond, he comes up with a unique equation. (Bob, student 240097, Plate tectonics listening-writing task)

Such correspondences were also evident for individual students across the various task types. For instance, Tracy consistently observed that the performances of student 130005 on several of the prototype tasks matched closely what she had seen him do in tasks in her class:

> That is a classic example. That is bang on the way he always speaks in class and out of class. Perhaps a tad slower. He will often speak very quickly. He doesn't hesitate much unless he's thinking of something to say. There was for me, or for a nonarchitect, some vagueness in what he was referring to. It makes perfect sense if you know architecture and deconstructivism, but if you don't, he doesn't clarify, but he would do the same thing in his first language. So that's exactly the way he speaks. It's a perfect sample. (Tracy, student 130005, Independent speaking task)

> Absolutely typical of this person. There's the attempt to organize things with "first of all" and "second of all" and "finally" so they know what's expected in terms of this kind of piece of writing. Very close to what they do in any kind of timed example in class, the revisions to little translations from the original language, the common spelling errors, or verb tense errors that come up every time the person is under stress. Same brand of mistakes, a little bit of Spanish word order. So this is identical to what the person produces. (Tracy, student 130005, Dance reading-writing task)

Particularly interesting were instances where the instructors thought the students performed better on the prototype test tasks than they usually did in their classroom activities. For example:

> This is better than the student has usually done. I think this is in fact more clever. It's, it's also more than, than he has usually, you know this is probably the most I've seen from him in a timed, in a timed writing piece. This is, this is a good, uh, a good piece to come out. (Tyler, student 130055, Dance reading-writing task)

> This is some of the student's better work, where they've been thrown into an unfamiliar situation and have used the discipline—that is reflective of this student as a learner—he remembered to apply the strategies irrespective of the level of difficulty. So, one of the more successful products of this person, who I think performs well under pressure. (Tracy, student 130010, Dance reading-writing task)

But, as indicated above, the instructors also thought that about one in five of the students' performances on the prototype tasks were worse than what they usually did in their ESL classes. The instructors often suggested that the reason appeared to be the students' lack of comprehension of the stimulus reading or listening materials. For instance:

> She's done less well here than she has in some other academic speaking tasks in trying to describe information. She seems to have a lot of trouble connecting the ideas. She's got some isolated details, but the way she presents them is so disjointed that it's not clear that she's understood the relationship between the ideas or exactly what their significance is. Uh, you hear a lot of hesitation, more than normal, where she may be searching to remember something, or searching for the right way to communicate that, but the way she's performing here, it's not clear whether it is a lack of comprehension or just searching for the right way to present the information. I suspect that she's missed a lot of this. She stays with the very basic ideas. Even with the numbers, she said, 4 years 5 months, but it's not clear to me that she knows what the significance of those ages is. So not as successful here as she has been. (Tracy, student 130011, Innate vs. learned behavior reading-speaking task)

Uh-oh, ok, so he was reading directly from the text; he chose the part about the study, but then he connected it into the other study, so he really was not able to complete the task, uh, probably based on the fact that he could not understand the reading. The content was unfamiliar; his reading is probably fairly slow. Uh, probably he did not understand a lot of vocabulary. (Ann, student 240009, Innate vs. learned behavior reading-speaking task)

Ok, she did not do too well on the response here and my suspicion is that it is due to her listening. I do not think she understood the lecture. I do not think she would have the problem with these ideas, for example, if it was written out, she could probably respond. I suspect the problem was her listening, so she really did not answer anything clearly or correctly—so I am not surprised one bit if it is the listening. (Steve, student 24028, Plate tectonics listening-writing task)

Students' limited comprehension of the content of the stimulus material also appeared to combine with limitations in their vocabulary or lack of familiarity with the topic or genre:

I think it's poorer than we usually saw…I think it has to do with, again, as with the last case, the complexity of the material. She was really struggling to find the words to describe the experiment. And she really only got the first half of it…You know at one point there she was trying to say that, the, uh, they "discriminate," and she stumbled over the word about three times, and before trying to find an alternative way of phrasing that point. So her real problem here is the nature of the material, and the lack of vocabulary to express it, to articulate it, so that's a little poorer. Nothing in our class brought her up against material at that level. (Francis, student 130070, Innate vs. learned behavior reading-speaking task)

Uh, she did not exactly explain the three types correctly. I think I would say it is not the greatest sample of her writing, it is not as good as she could do. I do not know why, maybe, I guess—the topic—I would guess is not familiar to her at all. Uh, because I think her listening ability is probably pretty good, maybe she was not able to get all of the details down about the three types, so then she was not able to write about them well. I

think that is probably what happened. (Ann, student 240022, Plate tectonics listening-writing task)

Uh, I would say that was not the greatest. She was reading directly from the text, so that didn't demonstrate her ability to explain the passage. Plus her pronunciation was really bad. Uh, I think actually that the reading passage itself was so hard that she was not even able to understand most of what it was talking about. Uh, based on what she did in my classes, uh, various readings that we did, she always did really well with those, and so I think this text was probably completely unfamiliar—the topic, the content, was so unfamiliar to her that she just, you know, could not even paraphrase it herself, so she read directly from the text. The encouraging thing is that she chose the right part of the reading to copy, but she was not able to say it in her own words. (Ann, student 240065, Innate vs. learned behavior, reading-speaking task)

But limitations in comprehension, vocabulary, and background knowledge were also evident in some students' performances that the instructors' thought matched those they usually saw in classes. For instance:

What she has written shows only partial understanding of the difference between the two types of dance. She touches on the topic of each difference, but very superficially in some places in some places, she misunderstands what was the original meaning. I think it is kind of difficult from the vocabulary (in the reading), and that is probably typical of what she is capable of doing. Maybe I expected a little bit better, but this is about what she can do. (Steve, student 240029, Dance reading-writing task)

This is the written response explaining how modern dance is different from classical ballet. And the sample probably represents his ability. I think his reading ability—he may be fairly slow. The topic was probably completely unfamiliar to him. The writing he did in this sample, he is copying directly from the reading, so he did not really explain the difference between the two types and he did not do it in his own words, really, so it's not a very good writing sample. And probably it is based on his reading ability, which is probably fairly slow. (Ann, student 240066, Dance reading-writing task)

In addition, some students may have lacked practice with the task conditions. For instance, Sara's students had not practiced timed writing in her class, so Sara thought this affected their performance on the prototype tasks:

> A lot of the writing I've seen was not writing done in a certain time limit, so I would say it seems a bit better in class than this is. It seems that some of the things are the same, like she's trying to use some organizational words, and so on, which she did, and some of the mistakes are the same also. And some of the correct use of "for example," of articles, were the same as what she does, what she did in class. But I would have expected her, let's see, yeah, I would expect her to do a little bit better in class. (Sara, student 130013, Dance reading-writing task)

Others may have been accustomed to handling readings in specific ways:

> I've noticed that when the students read, they have a marker and they use a marker to highlight their reading. And not only do they use a marker, but they use the edge of the reading material, and they write comments. So this might be something that is, perhaps, their first experience to listen and then respond that way. (Bob, student 240102, Plate tectonics, listening-writing task)

Or other students may not have grasped the unique nature of the task or the prompt for it:

> Well, she did not answer the prompt; she misunderstood it or did not read it completely, I think. I think she may have read the first sentence of it or something, so I would say that the sample did not represent her ability. Uh, what it did represent—her ability, her pronunciation, her fluency, but, uh, as far as doing the task, it was not a good representation of her ability, I would say. I have heard her do better than that in class. (Ann, student 240026, Independent speaking task)

Certain students just seemed to have floundered, exacerbating their performance in multiple ways:

> I think that the Dance reading was somewhat of a disaster for this person because they weren't familiar with the topic and I think they were paralyzed with the unfamiliarity, and that's common for this person. If they feel comfortable, they're okay, but if they realize they're in foreign territory, they freeze up. I'm struck here by the fact that this person could have written about her own experiences, and therefore had showed common sense in choosing what to write about, and yet it's very basic language, basic language there. (Tracy, student 13008, Dance reading-writing task)

Research Question 3: Fulfillment of Evidence Claims for the Prototype Tasks

Two types of information were solicited to address this research question. First, the seven instructors provided, in their questionnaire responses, impressionistic evaluations of the prototype tasks in view of the constructs that each task was intended to realize as well as the performances that their students had demonstrated on these tasks. Second, we correlated the independent ratings that the instructors had given of their students' abilities to perform the kinds of speaking and writing that were supposed to be represented in these tasks with the scores that the students were given on these tasks in the field tests.

Writing tasks. The instructors' evaluations of the independent writing task (i.e., TOEFL essay about personal experiences and opinions) ranged from "very effective" (ratings of 4 or 5 from four instructors) to middling (ratings of 3 from three instructors). The flexible opportunities this task offered for personal expression were mostly cited as its chief virtue. Tracy was positive about this task because "students have freedom to explore a wide variety of experiences they've discussed and/or written about before." Tyler similarly observed how students were "able to respond to situations with which they were familiar." Bob thought the "open-ended prompts lead to individual interpretations." Steve noted that "everyone was able to write about these topics." Ann stated, "The topics to choose from were general enough to enable various proficiency levels of students to write their ideas."

Francis expressed a contrary view, remarking that "many students are very reluctant, for personal and/or cultural reasons, to freely express themselves in this area," observing that the effectiveness of the independent writing task was variable: "it worked better for some of our students than for others." Sara contrasted this writing task with the other ones that involved reading or listening, claiming she thought the TOEFL essay was a "better test of their writing"

because the students "didn't need to understand a reading/listening" passage prior to writing. At the same time, Sara criticized the time limit on this essay task and called it "formulaic." (Note, however, that Sara's students had a lower level of proficiency in English than the students taught by the other instructors in Toronto, and this may have influenced Sara's views.)

The instructors' evaluations of the prototype writing tasks that involved responses to reading or listening materials were similarly positive (ratings of 4 or 5 on the scale of 5 for four instructors) to middling (ratings of 2 or 3 for three instructors). Six instructors praised the authenticity of these tasks, particularly for their correspondence to real texts and performance in academic settings. For example, "The lectures were good simulations of real academic situations"; "Typical of material they would encounter in university settings"; "seems to be a true academic task"; "it is an authentic type task (one that would be expected in an academic class)"; "lecturers paraphrased and gave examples in an authentic way"; and "they dealt with analysis and theory as academic texts do."

Sara, however, judged these tasks less positively, observing that her students' writing performance was heavily dependent on their comprehension of the reading or listening source materials they had to write about. As she put it, "They had to understand the lecture. It's very hard to write clearly when you don't know what you're writing about." The limitations the other instructors observed about these tasks focused on two concerns. One concern was about bias or unfamiliarity in the topics: "previous experience and background cultural knowledge (especially Western culture) would predispose a student to do well"; "Asians may not be familiar with ballet as an art form—the topic seems to be somewhat culturally biased"; "outside of area of interest"; "the lecture topic was perhaps unfamiliar to many students"; "choices of different topics are recommended within that framework of delivery"; "I think that more business-oriented or social science/education/public administration topics should be considered for readings and lectures. Many students starting graduate school in the U.S. are going to be studying in these areas (more so than dance or plate tectonics type subjects)." The other concern related to the presentation of the stimulus materials: "they spoke slowly and paused much more than lecturers at university tend to"; "the lectures were very short." In the interests of facilitating examinees' comprehension of the stimulus materials, Steve recommended "an outline or other visual aids," and Ann suggested that "PowerPoint-type slides (nonmoving visuals) or diagrams/pictures would be useful visuals."

*Speaking tasks***.** The instructors rated the prototype speaking tasks positively overall (ratings of 4 or 5), except for Sara, who rated the speaking tasks in response to readings or lectures as limited in their effectiveness (giving these ratings of 2). As with the comparable writing tasks, Sara felt her students did not perform well if they had not fully understood the stimulus materials they had to speak about.

The independent speaking task was rated positively by all the instructors (all rating it 5 out of 5 except for Francis, who rated it 3). They praised its "straightforward questions and situation"; "freedom to explore topics of interest"; "a topic that most students could speak on"; "real life tasks"; "many opportunities for students to recall their past experiences and to use their basic communicative skills"; and the "clear sample" of students' speech obtained. These comments resembled those that the instructors gave to the independent writing task. Limitations the instructors observed in this task related to its time constraints and the affective states of students speaking independently into a microphone ("anxiety can be a barrier to effective performance"; "some students won't/can't respond to this sort of task"; "seemed too short; not time for development") or the formulaic nature of the students' speech ("they may have given this speech before in ESL classes"; doesn't necessarily elicit abstract ideas").

Some of the instructors evaluated more positively the speaking tasks that involved responding to listening materials (ratings of 4 or 5 for all instructors) than the tasks that involved speaking about written texts (which three people rated as 2 or 3 out of 5). They praised these speaking tasks for the opportunities they presented to students: "forces students to speak about abstract topics"; "being put on the spot to orally explain what one has heard in class is a skill students require and one which they often lack"; "students encounter such situations in classrooms and need more experience and exposure." Likewise, the instructors praised the realism of the stimulus texts, particularly the lecture ("the lectures balanced theory and abstract thought with examples and concrete detail in a realistic mix"; "lecture was clear; the topic not too specialized"; "visuals, vocabulary, geographical location were identified to reinforce the oral presentation") and the conversation ("very typical conversation–realistic exchanges"; "idioms and natural miscommunications or vagueness were included in the conversations"), and in some instances also the reading stimuli ("the range of texts used was revealing—pointed towards the importance of contextual understanding in reading comprehension"; "good task, but...").

The limitations the instructors observed in these tasks likewise concerned qualities of the stimulus materials: "the passages were dense; students may have had trouble remembering information under pressure"; "the speakers spoke slowly and repeated ideas more frequently than normal. There was no background noise"; "the conversation-based task used with our students was not academic in its nature/topic"; "Most but not all students here are familiar with the background. If not, would it go over their heads?" Some instructors also queried the abilities that students demonstrated in these integrated tasks: "memorization and repetition of the lecture passages may have taken place"; "Asian speakers…were less confident"; "wide range of skills and abilities in the class and therefore difficult to get [an] accurate reflection" [of their abilities]." This criticism focused particularly on the task that required speaking about a reading text: "Reading text and topic were tough. I noticed only the very high-level students handled it as well as their ability really was. Weaker students didn't seem to stand a chance."; "Too technical? Many students just quoted parts of the text."

Instructors' ratings and scores on the field test. The instructors' ratings of the students' abilities to speak about personal experiences and opinions correlated significantly, showing a large effect size, with the students' scores on the independent speaking task: *rho* = .48 ($p < .01$), $r = 1.1$, $n = 28$. As shown in Table 5, none of the other ratings of the students' abilities by the instructors correlated significantly with the students' scores on the prototype tasks that were designed to assess corresponding constructs. The only other correlation to be statistically significant and to show a large effect size was the correlation between the two writing tasks that involved writing from sources (i.e., the listening-writing task about plate tectonics and the reading-writing task about dance): *rho* = .50 ($p < .01$), $r = 1.2$, $n = 41$. It should be noted, of course, that because individual test items or tasks are typically of marginal reliability (when compared to the reliability of a complete test), the correlations of individual task performance with instructors' ratings can be expected to be relatively low. However, some of these correlations did show small effect sizes, indicating that a larger sample of teachers and students might have produced significant correlations among the variables.

Table 5

Spearman Correlations Between Instructors' Ratings of Individual Students' Abilities in English and Scores the Students Received on the Prototype Tasks Intended to Assess These Abilities

	Speak independently	Speak about personal topics	Speak about academic topics[a]	Write independently[b]	Write about personal topics	Write about academic topics
rho	.48 ($p < .01$)	.25	.09	.09	.09	.09
r	1.1	.44	.28	.18	.18	.34
n	28	26	26	41	41	41

Note. rho = Spearman correlation, *r* = effect size, *n* = number of student performances evaluated; missing data were not included in the correlations.

[a] The test score for speaking was in response to a lecture on ground water. [b] The test score for speaking was in response to a conversation about student housing.

Discussion

Although exploratory and small in scale, this study provides considerable qualitative information that demonstrates the educational relevance, authenticity, and content validity of the prototype tasks being field-tested for a new version of the TOEFL test. At the same time, these findings point toward revisions that might usefully be undertaken on certain aspects of the prototype tasks being field tested for the new TOEFL.

Considering tasks that (a) involve independent writing or speaking and (b) involve writing or speaking in response to reading or listening materials, seven highly experienced ESL instructors expressed positive impressions of the prototype tasks, judging them to be realistic and appropriate in their simulation of academic content and situations, in the skills they required students to perform, and in the opportunities they provided for students to demonstrate their abilities in English. Importantly, the instructors thought the prototype tasks permitted the majority of their students to perform in English in the test contexts in ways that corresponded closely to the performance those students usually demonstrated in their ESL classes and course assignments. As such, the present results contribute some limited evidence for the construct validity of these prototype tasks, suggesting that the prototype tasks field-tested here have the

potential for positive washback on educational practices. Nonetheless, the value of this research is primarily in verifying that experienced ESL instructors consider the prototype tasks to be performing as intended. In instances where certain tasks were not, the findings suggest ways in which these tasks might be improved in further field trials.

The instructors' impressions of the prototype tasks and of their students' performances on them support the inclusion on the new TOEFL of the two principal types of tasks field-tested here: (a) independent writing and speaking tasks that permit students to choose how they express themselves in English, and (b) tasks that integrate students' reading of or listening to specific academic texts then either writing or speaking about them. The instructors had few criticisms to make of the independent writing or speaking tasks, apart from observing that they tended to invite rhetorically formulaic kinds of language production. Although familiar with the independent writing task (as the TOEFL essay or the Test of Written English), the instructors appeared to welcome the parallel version of it in the form of the independent speaking task, and to find their students performed well on it. The instructors praised the new integrated tasks for their authenticity with respect to the demands of academic studies in English, but they also raised numerous criticisms that suggest how such tasks might be optimally designed.

The most prevalent problem observed in the integrated tasks was that students with lower proficiency in English were hampered in their speaking or writing performance if they did not comprehend the ideas, vocabulary, or background context of the reading or listening stimulus texts. This problem of task dependencies, which was anticipated from the outset of task development, has commonly been observed as a limitation in language tests that use interpretive texts as the basis for language production tasks (e.g., Clapham, 1996). To this end, several instructors suggested that topics be selected for such texts that most students would be fairly familiar with and interested in; visual or schematic materials be appended to guide examinees' interpretations of the principal content in and rhetorical organization of the texts (cf. prototype tasks analyzed by Ginther, 2001); and examinees need to be familiar with and practiced in the task types and their prompts. At the same time, the instructors praised characteristics of the prototype tasks—such as their authenticity, natural language, and academic orientation– suggesting these characteristics be retained as primary features of the tasks. Indeed, some instructors thought the tasks might be made even more challenging in terms of their cognitive or intellectual demands. In turn, many of the instructors' remarks about their students' performance

on the prototype tasks indicate that the limitations that some students experienced in comprehending the content, vocabulary, or significance of the stimulus materials realistically reflected limitations in their abilities in English. Given that the sample of ESL students here included a range of students who were studying English just prior to attending, or who had just been admitted into their first year of, universities in the United States or Canada, this tendency indicates that the present prototype tasks may be pitched at a level that might demarcate distinctly between students who do have, or do not yet have, the requisite proficiency in English for academic studies. Students lacking abilities to read or listen to academic texts then write or speak about them in English might simply not be able to fare well in this new version of the test until they have acquired such abilities. Orientation or other educational materials could help such students determine their readiness for the test and to prepare themselves for it.

Interpreting the correlations between the instructors' ratings of their students' abilities and the students' performance on the prototype tasks is difficult for several reasons: Only a small number of teachers (and small samples of their respective students) participated in the study; over half the data from speaking tasks were lost in transmission from the field test sites; and the number of tasks that could reasonably be considered in an interview was too small to be able to distinguish between the construct of the task and the topics and prompts through which each task was realized. Moreover, as mentioned earlier, because individual tasks can be expected to be of only modest reliability when compared with the reliability of a complete test, correlations were necessarily restricted. We were therefore not able to demonstrate, as evidence for construct validation, that scores on the prototype tasks corresponded to teachers' judgments about relevant aspects of their students' abilities. Nonetheless, many suggestive directions arise from these results.

The most striking quantitative result was that scores on the independent speaking task correlated with the instructors' ratings of their students' usual abilities to speak about personal experiences and opinions. This finding seems to verify the worth and relevance of this task type, which has not previously appeared on the TOEFL test. Likewise, the correlations between the two tasks that required students to write about either reading or listening passages suggest these tasks may be assessing similar constructs. But the lack of correlations between the instructors' ratings of their students' abilities and the students' scores on the prototype tasks is puzzling (apart from the empirical constraints outlined in the preceding paragraph).

One possible interpretation of this pattern of findings is Cummins's hypothesis about teachers' abilities to distinguish between their students' basic communicative interpersonal skills (BICS) and cognitive-academic language proficiency (CALP) (e.g., articulated in Cummins, 1984, and refined in subsequent publications). Cummins's research suggests that ESL teachers are able, from frequently observing their students' oral performance in classes, to make relatively accurate judgments about their students' BICS, but that they seldom have access to information about their ESL students' CALP because that is usually exercised mentally and privately in the context of tests, reading, or studying, so teachers have difficulties evaluating these abilities. This tendency may have applied to the present instructors as well. But given their focus on teaching writing and academic skills to their students, and their claims to have known their students' abilities in these domains very well, this is a speculative interpretation at best. Another possible interpretation is that the prototype writing tasks differed from the types of writing the students usually did in their courses (i.e., the prototype tasks were single drafts written under strict time constraints, rather than multiple-draft compositions revised at students' leisure between classes), as Sara remarked in her interviews, so the instructors' ratings may have referred to different conditions for writing than in a test context. In this regard, regular university professors, rather than ESL instructors, may be informative judges of the content validity of prototype test tasks, as demonstrated by Elder (1993) or for the new TOEFL by Rosenfeld et al. (2001). Moreover, the instructors themselves may have differed from each other in their standards for assessing students' writing because they each taught different courses and worked at different institutions. Alternative explanations may well be considered within the scoring schemes for the prototype tasks and the nature of the tasks themselves, which would appear to warrant refinements, particularly the three tasks (i.e., on plate tectonics, dance, and innate vs. learned behavior) that elicited numerous performances that the instructors judged to be worse than their students usually did in classes. Such tasks did not appear to have a high-level schematic organization to their presentation of information that students could easily grasp in their comprehension or convey in their writing or speaking, suggesting that a basic level of rhetorical, schematic organization of information might be required in reading or listening source texts to make them viable for integrated tasks on a test like the TOEFL.

References

Aaron, B., Kromrey, J.D., & Ferron, J.M. (1998, November). *Equating r-based and d-based effect size indices: Problems with the commonly recommended formula*. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED 433 353).

Alderson, J.C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *TESOL Quarterly, 13*, 280-297.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Bailey, K. (1999). *Washback in language testing* (TOEFL Monograph Series, Rep. No. 15). Princeton, NJ: ETS.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series, Rep. No. 19). Princeton, NJ: ETS.

Bridgeman, B., Cline, F., & Powers, D. (2002, April). *Evaluating new tasks for TOEFL: Relationships to external criteria*. Paper presented at the annual meeting of Teachers of English to Speakers of Other Languages (TESOL), Salt Lake City, UT.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programs: A review of the issues. *Language Testing, 15*(1), 45-85.

Brindley, G. (Ed.). (2000). *Studies in immigrant English assessment* (Vol. 1). Sydney, Australia: Macquarie University, National Centre for English Language Teaching and Research.

Butler, F., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series, Rep. No. 20). Princeton, NJ: ETS.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Chapelle, C., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000* (TOEFL Monograph Series, Rep. No. 10). Princeton, NJ: ETS.

Clapham, C. (1996). *The development of the IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, England: Cambridge University Press.

Cumming, A. (1996). The concept of validation in language testing. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 1-14). Clevedon, England: Multingual Matters.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series, Rep. No. 18). Princeton, NJ: ETS.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, England: Multilingual Matters.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5-30.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*, 233-254.

Enright, M., & Cline, F. (2002, April). *Evaluating new task types for TOEFL: Relationship between skills*. Paper presented at the annual meeting of Teachers of English to Speakers of Other Languages, Salt Lake City, UT.

Enright, M., Grabe, B., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series, Rep. No. 17). Princeton, NJ: ETS.

Epp, L., & Stawychny. (2001). Using the Canadian Language Benchmarks (CLB) to benchmark college programs/courses and language proficiency tests. *TESL Canada Journal, 18*(2), 32-47.

ETS. (2001). *New TOEFL: Phase 3 prototyping interim report*. Unpublished manuscript.

Freedman, S. (1991). *Evaluating writing: Linking large-scale testing and classroom assessment.* (Occasional Paper 27). University of California, Berkeley: Center for the Study of Writing.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics, 20*(2), 221-236.

Ginther, A. (2001). *Effects of the presence and absence of visuals on performance on TOEFL CBT listening-comprehensive stimuli* (TOEFL Research Rep. No. 66). Princeton, NJ: ETS.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Grant, L. (1997). Testing the language proficiency of bilingual teachers: Arizona's Spanish proficiency test. *Language Testing, 14*(1), 23-46.

Griffin, P. (1995). *The American literacy profile scales: A framework for authentic assessment*. Portsmouth, NH: Heinemann.

Hamp-Lyons, L,. & Kroll, B. (1997). *TOEFL 2000–writing: Composition, community, and assessment* (TOEFL Monograph Series, Rep. No. 5). Princeton, NJ: ETS.

Hoge, R., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297-313.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph Series, Rep. No. 16). Princeton, NJ: ETS.

Kunnan, A. (1998). Approaches to validation in language assessment. In A. Kunnan (Ed.), *Validation in language assessment* (pp.1-18). Mahwah, NJ: Erlbaum.

Lewkowicz, A. (2000). Authenticity in language testing. *Language Testing, 17*, 43-64.

Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Miesels, S., Bickel, D. Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgements: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38*(1), 73-95.

Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445-465.

North, B. (2000). *The development of a common framework scale of language proficiency*. Oxford: Peter Lang.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly, 24*(3), 427-442.

Rosenfeld, M., Leung, S., & Oltman, P. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph Rep. No. 21). Princeton, NJ: ETS.

Sharpley, C., & Edgar, E. (1986). Teachers' ratings vs. standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23*, 106-111.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing, 18*, 463-481.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2*, 31-40.

Stansfield, C., & Kenyon, D. (1996) Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 124-153). Clevedon, England: Multilingual Matters.

Sullivan, B., Weir, C., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

Teachers of English to Speakers of Other Languages. (1998). *Managing the assessment process: A framework for measuring student attainment of the ESL standards*. Alexandra, VA: Author.

Thompson, B. (2000). *A suggested revision to the forthcoming 5th edition of the APA Publication Manual.* Retrieved September 5, 2002, from http://www.coe.tamu.edu/~bthompson/apaeffec.htm

Notes

[1] We also gathered supplementary data from two instructors at the University of Bridgeport and two instructors at Central Michigan University, who had from 1 to 3 years of ESL teaching experience. Initial analyses showed that the responses of these less experienced teachers were less well informed than those of the seven highly experienced instructors, so data from the former are not included in this report.

[2] Aaron, Kromrey, and Ferron's (1998) formula produces an index of effect size (r) accounting for variance (rather than just standardizing mean differences) through the formula:

$$r = d / [(d2 + [N2 - 2N) / (n1\ n2)]].5 \text{ where } N = n1 + n2$$

This indicator of effect size is appropriate for nonexperimental studies when "total sample size is small or group sizes are disparate" (Thompson, 2000, p. 2). The interpretation of r is that 0 is a trivial effect, 0.2 is a small effect, 0.6 is a moderate effect, 1.2 is a large effect, 2.0 is a very large effect, and 4.0 is a nearly perfect effect.

Appendix

Profile Questionnaire

The purpose of this questionnaire is to gather background information about you and your overall impressions of the prototype tasks for the project, *A Teacher-Verification Study of TOEFL Prototype Tasks.* As with other data generated for this project, your identity will remain confidential.

The pseudonym I would like to use is: _____

Date of completing this questionnaire: _____

I. The Prototype Tasks

1. Overall, how well do you think the prototype tasks elicited the performance of the students from your class? For each skill, please circle the number that best corresponds to your answer. Circle 0 if you can't tell.

	Not Effective				Very Effective	
Writing	1	2	3	4	5	0
A. About personal experiences and opinions	1	2	3	4	5	0
B. About academic topics						
1. in response to lectures	1	2	3	4	5	0
2. in response to reading text	1	2	3	4	5	0
Speaking	1	2	3	4	5	0
A. About personal experiences and opinions	1	2	3	4	5	0
B. About academic topics						
1. in response to lectures	1	2	3	4	5	0
2. in response to conversations	1	2	3	4	5	0
3. in response to reading text	1	2	3	4	5	0

2. For each of the tasks, please indicate the ways, if any, in which they were effective and also the ways, if any, in which they were not effective. If there were any specific tasks that you felt were effective or not effective, please describe which tasks.

Writing

A. About personal experiences and opinions

Effective because _____

Not effective because _____

B. About academic topics

1. In response to lectures

Effective because _____

Not effective because _____

2. In response to reading text

Effective because _____

Not effective because _____

Speaking

A. About personal experiences and opinions

Effective because _____

Not effective because _____

B. About academic topics

1. In response to lectures

Effective because _____

Not effective because _____

2. In response to conversations

Effective because _____

Not effective because _____

3. In response to reading text

Effective because _____

Not effective because _____

3. Overall, how well do you think the tasks represent the domain of academic English required for studies at an English-medium university?

Not well				Very well
1	2	3	4	5

Why? _____

_____

Which tasks do this best? _____

_____

What do you think is missing? _____

_____

II. Personal Profile

1. My gender is: Male ____ Female ____

2. My age is: ____

3. I have taught English for _____ years.

III. Current Teaching Situation

1. My current role is: Assessor ____ Instructor ____ Administrator ____

Student ____Researcher _____ Other (specify)_____

2. The context(s) I have mostly worked in is: English (mother tongue) ____ESL ___

EFL ____ ESP ____Other (specify) _____

3. I have taught the present students for _____ weeks.

4. I know their abilities in English:

	Not well				Very well
Writing	1	2	3	4	5
Speaking	1	2	3	4	5
Reading	1	2	3	4	5
Listening	1	2	3	4	5

IV. Language(s)

1. My first language is: _____

2. My dominant language at home at present is: _____

3. My dominant language at the workplace is: _____

V. Educational History

Please describe your educational background in terms of:

	Degree/diploma/ certificate	Subject area	Language of education
1. Undergraduate studies			
2. Postgraduate studies			
3. Professional certification			
4. Any specialized training related to assessment			

VI. Experiences Teaching and/or Assessing English

1. Please describe any significant teaching and/or assessment experiences that might have influenced your assessments of the ESL students' abilities in the present research project.

_____

_____

_____

2. How would you describe your own skill in assessing English as a second language?

_____ Expert _____ Competent _____ Novice

3. How many years' experience do you have in assessing ESL? _____

4. Have you given any training courses in assessing language performance or administered such programs? If so, please describe these briefly.

_____

_____

Thank you for this information!

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

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

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

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

Email: toefl@ets.org

Web site: www.ets.org/toefl

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

I.N. 998777