ETS TOEFL

Monograph Series

MS - 30
April 2006

Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for the New TOEFL

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

Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for the New TOEFL®

Alister Cumming
Ontario Institute for Studies in Education of the University of Toronto, Canada

Robert Kantor
ETS, Princeton, NJ

Kyoko Baba, Keanre Eouanzoui, Usman Erdosy, and Mark James
Ontario Institute for Studies in Education of the University of Toronto, Canada

RM-05-13

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 TOEFL Internet-based test (TOEFL iBT), 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 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 TOEFL iBT, which was introduced in September 2005.

TOEFL Program
ETS

Abstract

We assessed whether and how the discourse written for prototype integrated tasks (involving writing in response to print or audio source texts) field tested for the new TOEFL® differs from the discourse written for independent essays (i.e., the TOEFL essay). We selected 216 compositions written for 6 tasks by 36 examinees in a field test—representing Score Levels 3, 4, and 5 on the TOEFL essay—then coded the texts for lexical and syntactic complexity, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source text. Analyses with nonparametric MANOVAs, following a 3-by-3 (task type by English proficiency level) within-subjects factorial design, showed that the discourse produced for the integrated writing tasks differed significantly at the lexical, syntactic, rhetorical, and pragmatic levels from the discourse produced in the independent essay on most of these variables. In certain analyses, these differences were also obtained across the 3 ESL proficiency levels.

Key words: Discourse analysis, ESL composition assessment, ESL proficiency, independent tasks integrated tasks, second-language writing, text characteristics

Acknowledgments

We thank Don Banh of the Education Commons at the Ontario Institute for Studies in Education for designing the computer software program for analyses of verbatim strings of words from source texts, type-token ratios of lexical words, and word length.

Table of Contents

List of Tables

Page

Introduction

Purpose

This report describes a study addressing the question posed for the research agenda for the new TOEFL®: How does the written discourse generated by independent tasks differ from the discourse generated by integrated tasks? The context of this research was the field testing for the new TOEFL of prototype task types that involve examinees writing to integrate reading and listening stimulus materials (i.e., integrated tasks) in addition to writing compositions for the essay task that is now featured in the TOEFL computer-based testing (CBT) as the TOEFL essay (i.e., an independent task, formerly the Test of Written English™, which does not require references to source materials in the exam materials). The rationale for the research study stems from numerous arguments and sources of evidence that have suggested that adding these new types of writing tasks to the TOEFL will diversify and improve the measurement of examinees' writing abilities, improve the washback effects of the test on teaching and learning practices internationally, and require examinees to write for the test in ways that more authentically resemble the types of performance needed for academic studies at universities in North America (Cumming, Grant, Mulcahy-Ernt, & Powers, 2004; Cumming, Kantor, & Powers, 2001, 2002; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Hamp-Lyons & Kroll, 1997; Lee, Kantor, & Mollaun, 2002; Rosenfeld, Leung, & Oltman, 2001).

Various analyses have recently been undertaken at ETS to develop new integrated tasks (that combine reading, writing, listening, and speaking modalities in various combinations), to devise scoring rubrics for them, and to interpret the test score data emerging from field tests of prototype versions of them (e.g., as reported in the studies cited above, among others). These analyses have culminated in the publication of the *LanguEdge Courseware* (ETS, 2002), which describes and provides examples of these new task types, and provides practice test tasks and instructional materials, to help prepare examinees and English language educators for the new TOEFL when it becomes operational in the near future. But little systematic evidence is available to describe the actual features of the written discourse that examinees produce in these new tasks. Specifically, in what ways do the qualities of writing that examinees produce for these new integrated tasks differ from those they write for the existing independent essay on the TOEFL? Knowing that the writing of examinees differs across these task types, and knowing how they differ, is vital to validating these new writing tasks, particularly in justifying their value

on the test, for designing and interpreting scoring schemes for the writing component of the test, and for preparing orientation and training materials for the new TOEFL. A related issue concerns knowing if and how the written discourse may vary in the written compositions produced by examinees at different score levels on the test. The discourse of written texts cannot be assumed to be consistent for examinees with differing levels of proficiency in English, so consideration also needs to be given to how the written discourse of examinees varies in particular tasks with their English proficiency. This information is needed to verify, or refine, the scoring schemes being developed to evaluate examinees' performance on these writing tasks. Indeed, such information is necessary to realize the broad-based plans for validation of the new TOEFL in terms of the actual performance of examinees on the test, as initially envisioned in Jamieson, Jones, Kirsch, Mosenthal, and Taylor (2000) and as later proposed for the writing component in Cumming et al. (2000). As one effort to realize these goals, the present report documents a project we undertook from November 2002 to June 2003 to analyze the discourse features of compositions written by a purposive sample of English as a second language (ESL) students at three score levels who participated in a field test of prototype tasks for the new TOEFL test that was conducted in 2002.

Methods

The research involved three phases: (a) preliminary identification of relevant indicators of discourse features for analyses; (b) selection and coding of sample compositions; and (c) analyses of similarities and differences in examinees' performance across three types of writing tasks and across three score levels. In designing this study we opted for an approach to discourse analysis that involved relatively objective coding of characteristics of examinees' written compositions. This approach was intended to complement the more interpretive types of analyses that we and our colleagues had previously pursued in analyzing the prototype writing tasks, for example, based on the verbal reports of experienced raters performing holistic evaluations of examinees' essays in Cumming et al. (2001, 2002) or of experienced teachers judging the writing of their students on prototype tasks as in Cumming et al. (2004). Those previous studies had followed the premises of the *reader-writer* model for validating new task types for TOEFL (proposed in Cumming et al., 2000) by describing and evaluating the thinking and perceptions of experienced essay raters and of teachers as they read the writing produced for prototype tasks. The present study followed the premises of the *text characteristics* model—likewise discussed in

Cumming et al. (2000) and conventional to much previous research on the evaluation of second-language writing—in order to provide additional, complementary evidence about constructs inherent in the new task types and the scoring schemes for them.

Selection of Discourse Features

In Phase 1 of this research, we reviewed relevant published research on written discourse analysis and on ESL composition assessment, aiming to select a suitable range of discourse indicators for our project. We then discussed possible indicators critically among ourselves, trying them out on sample compositions (described below) to determine whether coding them was feasible and meaningful and could reliably be done on the present data. We decided early on that we wanted to select indicators that

1. would span a range of discourse features, including lexical, syntactical, rhetorical, and pragmatic characteristics—so our discourse analyses would comprehensively address each of these micro- and macro-level aspects of written texts, considered integral to ESL writing in academic contexts and important for the test, as described in Cumming et al. (2000) and Cumming (2001);
2. could be applied to each of the three task types (independent writing, reading-writing, listening-writing) equally well so as to be able to distinguish common and differing elements in them;
3. could be applied reliably and meaningfully (so we sought indicators that had been used in a range of previous studies and produced reliable, meaningful results, and that have clear theoretical justifications and operational definitions); and
4. might be expected to show differences between compositions scored at Levels 3, 4, and 5 so as to try to verify, or to refine if necessary, the scoring rubrics used to define these score points on the prototype scales for the new TOEFL, and thus also be relevant to the writing produced by adult ESL examinees who take the TOEFL.

After searching out (through library searches of recent journals and books) and reviewing a wide range of relevant publications with these criteria in mind, discussing potential indicators among ourselves over several weeks of meetings, and trying them out for feasibility of rating or coding, we decided on nine indicators to apply in our discourse analyses, building on precedents established in earlier studies of ESL writing and writing assessment:

1. *Text length*, operationalized as the total number of words written in a composition within the 30 minutes allocated for each task (Carlisle, 1989; Chenoweth & Hayes, 2001; Grant & Ginther, 2000; Homburg, 1984; Larsen-Freeman, 1978; Larsen-Freeman & Strom, 1977; Perkins, 1980; Polio, 1997; Reid, 1986; Way, Joiner, & Seaman, 2000; Wolfe-Quintero, Inagaki, & Kim,1998).

   We adopted the operational definitions appearing in Polio (1997).

2. *Lexical sophistication*, analyzed in two ways:
   a. *average word length* (Biber, 1986, 1988, 1995; Chipere, Malvern, Duran, & Richards, 2003; Engber, 1995; Frase, Faletti, Ginther, & Grant, 1999; Grant & Ginther, 2000; Reppen, 1994)

   Our operational definition (following Engber, 1995) appears in Appendix A.

   b. *type/token ratio* of the number of different lexical words over the total number of words per composition (Chipere et al., 2003; Cumming & Mellow, 1996; Engber, 1995; Grant & Ginther, 2000; Laufer, 1991; Laufer & Nation, 1995; Read, 2000)

   Our operational definition (following Engber, 1995) appears in Appendix B.

3. *Syntactic complexity*, analyzed in two ways:
   a. *number of clauses per T-unit* (Bardovi-Harlig & Bofman, 1989; Crowhurst & Piche, 1979; Faigley, 1980; Homburg, 1984; Ishikawa, 1995; Perkins, 1980; Polio, 1997; Stewart & Grobe, 1979;
   Wolfe-Quintero et al.,1998)
   b. *words per T-unit* (Chipere et al., 2003; Polio, 1997)

   Our operational definitions (following Polio, 1997) appear in Appendix C.

4. Holistic rating of *grammatical accuracy* as either 1 (many severe errors, often affecting comprehensibility), 2 (some errors but comprehensible to a reader), or 3 (few errors, and comprehensibility seldom obscured for a reader)

   Our operational definition of this scale (related to Hamp-Lyons & Henning, 1991) appears in Appendix D.

5. *Quality of argument structures*, evaluating the claims, data, warrants, proposition, oppositions, and responses to oppositions as six elements, each rated as either 0, 1, 2, or 3.

   Our operational definition of these scales (following Knudson, 1992; McCann, 1989; Toulmin, 1958; Toulmin, Rieke, & Janik, 1984; cf. Connor, 1990, 1991; Crammond, 1998; Yeh, 1998) appears as Appendix E.

6. *Orientations to source evidence* expressed in each T-unit, coding for *presentation of voice* (as references that are either unspecified or are specified with respect to evidence from self or from other(s), with a person either identified or not, or assume a common, communal knowledge) and for *functions and content of message* (as either a declaration, quotation, paraphrase, or summary)

   Our operational definitions of these categories (adapted and modified from Thompson, 1996; Plungian, 2001) appear as Appendix F.

7. Functional uses of phrases from source (a) reading prompts and (b) listening prompts that appear as *verbatim strings of words* in ESL compositions

   We hired a computer programmer in Toronto to prepare a software program that identifies all strings of three words or more that appear in examinees' compositions and that also appear in the source reading or listening materials for these integrated tasks in the field test. After using the computer program to tally these instances of verbatim phrases from the source texts, we reviewed all of our sample compositions to identify, and attempt to interpret, how examinees appeared to use the verbatim phrases from the source texts in their compositions. We then selected extreme cases (i.e., of ineffective, typical, and effective compositions) for each task type and score level to describe the functional purposes that these verbatim phrases appeared to serve for the ESL writers in their texts. In describing the discourse features of these case examples, we also exemplify the other discourse features treated quantitatively in our preceding analyses, aiming to describe how they function holistically for the range of compositions in our sample.

We did not make any predictions about how the indicators of *text length, lexical sophistication, syntactic complexity*, and *grammatical accuracy* might differ across the independent writing, reading-writing, or listening-writing tasks—or even within these task types or score levels. But we were predisposed to guess that the source texts for the integrated tasks might prompt examinees to create academically oriented contexts that could contain more sophisticated language forms than appear in the independent writing task (e.g., as found in Way et al., 2000), and to assume that compositions at higher score levels would demonstrate more proficient uses of lexical and grammatical indicators than would compositions at lower score levels (as established in the various publications cited above).

The indicator of *argument structure* catered to the independent writing task because examinees were instructed in this type of task to "formulate and convey in writing a response to a question that asks them to state and explain their position or opinion" on a particular topic, whereas the reading-writing and listening-writing tasks merely instructed examinees to answer a question about, and to "convey coherently, in writing, the relevant information" from, a passage they had read or a lecture they had listened to (see Appendixes G and H; cf. ETS*,* 2002, pp. 29-70; the operational definitions of these tasks quoted above appear on p. 29). We therefore expected that compositions written for the independent writing task would display more fully developed argument structures than would compositions written for the reading-writing or the listening-writing tasks (though these would necessarily involve some limited aspects of argument structure, as well). Likewise, we assumed that compositions scored at higher score levels (e.g., Level 5) would demonstrate more extensive argument structures than would compositions scored at lower score levels (e.g., Level 3).

In turn, we expected the indicator of *orientations to source evidence* to produce different results for each of the task types. In particular, we expected examinees to use information primarily from their own personal experiences or from their long-term memories to develop their ideas in the independent writing tasks (because they had no relevant source texts to reference). In contrast, we expected examinees to rely primarily on information from the source reading or listening tasks in the two integrated task types (as they were instructed to do). For these reasons we distinguished two aspects of these orientations to source evidence, based on Thompson's (e.g., 1996) extensive discourse analyses of oral language reports. First, we expected in the aspect of *presentation of voice* that compositions written for the independent task would involve

mostly references to the self as a source of evidence, whereas the integrated tasks would involve mostly references to others (i.e., the source texts) as sources of evidence. Second, we expected that the *functions of the message* would differ in each task type; in particular, we expected that the integrated writing tasks would yield compositions with references to information through quotations, paraphrases, or summaries of the source texts, whereas the independent writing task would produce more simple declarations without sources specified. We wondered if the oral versus written mode of presenting source materials might influence examinees' writing as well. Finally, our analyses of functional uses of verbatim strings of words from source texts focused only on the reading-writing and listening-writing tasks because there were no source texts for the independent writing task. Our interest here was to identify the extent of verbatim text borrowed directly from the source reading or listening texts and to interpret how examinees used such verbatim pieces of text in their compositions according to the mode of presentation of the source material (i.e., reading vs. listening) and to their levels of English proficiency (i.e., Score Levels 3 vs. 4 vs. 5).

It is also worth remarking on various types of discourse features that we did not select for our analyses, but which we did consider in Phase 1 of the project from our reading of previous, relevant research on summary or argumentative writing. We did not use approaches to discourse analysis that focused on specific categories of lexical choices because we wanted to use indicators that would span comprehensively the full range of writing that examinees produced in the present compositions, rather than just selected categories of words or phrases. For example, studies by Abdi (2002), Barton (1993), Biber and Finegan (1989), Conrad and Biber (2000), Crismore, Markkanen, and Steffensen (1993), Grant and Ginther (2000), Hyland (1996), and Ivanic and Camps (2001) have focused on coding the frequencies of specific word classes, such as adverbs, adjectives, or verbs, in texts that indicate functional purposes, such as hedging, emphatics, or markers of attitude or identity. Such analyses are certainly fruitful, but when we began to consider applying them to the present data, we found three reasons not to pursue them. First, we sought to focus our efforts on discourse features that correspond to the evidence claims guiding the design of the prototype tasks for the new TOEFL (i.e., as described in the *LanguEdge* materials as well as in Cumming et al., 2000). As noted above, we based our decisions to analyze argument structures and voice and message functions on this rationale, and to some extent these analyses subsumed other forms of functional or attitudinal expression that

we might have pursued (e.g., aspects of expressing identity are contained in our analyses of voice). In turn, these approaches led us in the direction of coding units of discourse (such as T-units, clauses, rhetorical discourse structures, and uses of evidence) rather than analyses of word classes, which would have been better addressed through computer-tagging programs, as have been featured in most of the studies cited immediately above.

Second, our initial reviews of the composition data determined that there were too few instances of explicit markers of concepts like individual identity or hedging, or that these were so ambiguously expressed in the compositions that they could not be interpreted reliably across the range of compositions in our sample to make analyses of such features worth undertaking here. A reason for this may be (we realized, upon inquiring into recent studies and theories of writing from source documents) that people's knowledge of relevant subject matter and of contextual situations exerts dominant influences on their evaluations and uses of source documents in their writing (e.g., Britt & Aglinskas, 2002; Rouet, Favart, Gaonach, & Lacroix, 1996; Stromso, Braten, & Samuelstuen, 2003; Wiley & Voss, 1999; Wineburg, 1994). In the context of writing for an exam like the new TOEFL; however, virtually all of these contextual factors have been eliminated (i.e., in the interest of making the content and conditions of the test fair to examinees with diverse backgrounds internationally; cf. Cumming, 2002). The context of a language proficiency test contrasts with the conditions for performance that might be expected for demonstrations of learning in academic courses, rendering it all but impossible for us to be able to evaluate evidence of contextual or knowledge influences in individual examinees' compositions. In turn, the discourse indicators that we chose are restricted to those that our preliminary analyses suggested could be coded reliably without extensive interpretation on the part of coders and without being susceptible to variation by task type or by the content or lexical knowledge expressed by individual examinees in their writing. The present discourse analyses are, therefore, characterized more by a parsimonious selection of stable text features than by their capacity to account fully for subtle nuances in the writing produced in each task type.

Third, we wanted to take an open-ended, neutral perspective on examinees' verbatim uses of phrases from the source reading or listening texts because we wanted to avoid prejudging how or why they may have used such phrases in their writing. As numerous authors have demonstrated, plagiarism can be difficult to evaluate in the writing of second-language learners (who are necessarily acquiring the language and thus appropriating it in various ways),

particularly outside of the context or explicit norms of a specific educational or institutional setting (Deckert, 1993; Howard, 1995; Kroll, 1988; Pennycook, 1996; Scollon, 1994; Shi, 2004). Likewise, we realized that our analyses could not meaningfully interpret examinees' composing strategies for making use of source texts in their compositions because we did not have data on individual examinees' composing processes, as have featured in many of the prior studies of writing from sources (e.g., Brown & Day, 1983; Cumming, Rebuffot, & Ledwell, 1989; Kintsch & van Dijk, 1978; McCarthy Young & Leinhardt, 1998).

Selection of Sample Compositions

We identified for analysis a purposive sample of 216 compositions written for 6 tasks in the 2002 field test of the new TOEFL test, as well as randomly selected 10 practice compositions (to try out our methods of analysis for feasibility, as described above) and 24 other compositions (to establish inter-coder reliability, as described below) from the pool of approximately 1,800 compositions written by about 300 students for this field test. For the field test, raters collapsed the Score Levels of 1 and 2 normally used for the TOEFL essay (or formerly, the Test of Written English) into a score level of just 1, because these TOEFL essay score levels could not be readily distinguished among the population writing the field test, and to make a 5-point scale that would be equivalent to the 5-point scale used for rating the integrated tasks in the field test (as shown in Appendix G). Scores we report below as 3, 4, and 5 on the field test would, for these reasons, correspond to scores of 4, 5, and 6 on the TOEFL essay. We did not have direct access to information about the individual characteristics of the examinees who had written the compositions other than to know that they were either in pre-university or university-credit ESL courses at a variety of universities in North America, Australia, and Hong Kong that had participated in the field test. The raters of the essays were experienced evaluators on staff at ETS who regularly scored compositions for the TOEFL essay.

To select the compositions, we first identified all examinees whose compositions were given the same score by two raters for the two independent essays in the field test. Knowing that the independent essay task and its scoring scheme have been implemented for many years as the Test of Written English (TWE®) and the TOEFL essay, we considered its scores to be a stable criterion for selecting our sample of compositions (cf. Stansfield & Ross, 1988). Moreover, by selecting only compositions that had been given the same score by two raters we presumed further stability in the score levels attributed to these writing samples. Only 12 compositions at

Level 5 met this criterion, so we set 12 as our sample size for this proficiency level, then we proceeded to select equal sample sizes for compositions scored at Proficiency Levels 3 and 4. That is, we sampled randomly (using a table of random numbers) another 12 compositions among those that had been given scores of 4 by two raters on the two independent essay tasks, and another 12 compositions that had been given scores of 3 by two raters on the two independent essay tasks. Writing scored at Levels 2 or 1 on the TWE scale did not appear to be amenable to the analyses we wished to do (because the writing was not sufficiently well-formed to be interpreted for its discourse features). That being the case, we did not sample from compositions scored at these levels. Moreover, Score Levels below 3 on the 2002 field test, or below 4 on the actual TWE, are below the proficiency levels usually considered for university admissions decisions, so we did not think that pursuing analyses with them would be useful in view of the overall purpose of the new TOEFL.

We next identified the compositions (which had already been scored by staff at ETS) that these 36 examinees wrote for two reading-writing and two listening-writing tasks in the 2002 field test. In this way, we gathered 216 compositions written by 12 people at Level 3, 12 people at Level 4, and 12 people at Level 5 (using the scores on the independent essays as the criterion of their score levels). The prompts used in the 2002 field test for these tasks have now been published in the *LanguEdge Courseware* (ETS, 2002), and we refer readers there for detailed descriptions of the task instructions, rating scales, and source texts for the integrated tasks (as well as to Appendixes G and H in the present report). Examinees had been given 30 minutes for each essay to write the independent essays, 25 minutes to write the integrated reading-writing tasks, and 15 minutes to write the integrated listening-writing tasks.

Table 1 shows the mean scores obtained by all examinees who wrote the six compositions in 2002 field test and the titles of the writing tasks. It is evident in Table 1 that the scores given to the independent essays tended to be about one point higher than the scores given to either the listening-writing or the reading-writing tasks. The compositions were conveyed from Princeton to Toronto then printed out and copied for the discourse analyses.

Procedures and Reliability for Coding, Rating, and Tallying Data

The discourse analyses were carried out by three members of our research team, all of whom were at the time doing a Ph.D. in second language education. Two of the researchers (Erdosy and James) were native speakers of Canadian English, and both had more than a decade of experience teaching and assessing composition in English as a second or foreign language, mostly in university contexts. The other researcher (Baba) had only a few years of EFL teaching experience, and English was her second language (Japanese being her first language). In addition to the perspective of a proficient nonnative user of English, Baba brought to the research team experience with various corpus linguistics projects from her previous studies.

Table 1

Task Types, Composition Topics, and Mean Scores for 300 Students in the 2002 Field Test

Task types and composition topics	Mean scores
Independent essays	
Independence	3.22
Plan a trip	3.17
Listening-writing	
Behaviorism	2.26
Ethics/Plato	2.54
Reading-writing	
Early cinema	2.26
Nineteenth-century politics in the United States	2.05

To establish inter-coder reliability before conducting the discourse analyses, we initially selected 24 compositions (4 from each of the 6 composition tasks), representing just over 11% of the total sample of 216 compositions. Three pairs of researchers, from the team in Toronto, did these preliminary analyses, working from printed copies of the compositions and the task instructions associated with them but blinded to the scores assigned to the compositions during the 2002 field test and blinded to the scoring rubrics contained in the *LanguEdge Courseware* (i.e., members of the research team who did the discourse analyses never saw the scoring rubrics

for the integrated tasks that appear in Appendix G until after the discourse analyses were completed). After initially establishing inter-coder reliability with other members of the research team, one member of the research team proceeded to do each of the specific analyses on the full data set, each person working on a particular analysis with which he or she was most familiar, experienced, and had proved reliable, and again blinded as to the scores of the compositions and the scoring rubrics for the integrated tasks (though each rater had, from limited past experiences, some familiarity with the scoring scheme for the TWE). As described above, the entire research team in Toronto had tried out each of the coding schemes in several half-day meetings prior to these reliability tests, and we established a consensus on operational definitions and problematic instances, refining our procedures for analyses initially during these sessions.

During the practice ratings of the 10 sample compositions, we modified slightly the coding categories and the operational definitions (e.g., we dropped a few initial coding categories for orientations to source evidence because we did not find any examples of them in the data). The coding categories and the operational definitions remained the same after that. We found that judging the quality of argument structures and orientations to source evidence involved more interpretations than did the other indicators, and it was difficult to reach a high inter-coder agreement on them (as described below). With that in mind, we developed a set of decision rules for coding these indicators based on the operational definitions, referring to concrete examples in the sample compositions (e.g., "if students write such and such, it should be coded as this"). Disagreements between coders were later resolved by discussion, and these were also reflected in the decision rules. We followed the decision rules when coding the 24 compositions to test inter-coder reliability as well as for coding the full set of data.

Segmentation of the compositions into T-units and clauses proved to be readily reliable, with two coders correlating at .99 (Spearman's *rho*) over the 24 sample compositions. Likewise, after we refined our operational definitions of coding categories, as described in Appendix F, the coding of voice and message in each T-unit was relatively consistent, producing percentage agreements of 94% for voice categories and 84% for the message categories (or Cohen's kappa of .75 and .72, respectively). Rating the subcategories of argument structure, however, was more challenging for two reasons. First, the various aspects of arguments were difficult to determine in the less proficient compositions. Second, some of the score points on the scales (adopted from Knudson, 1992, and McCann, 1989) were not used at all in the ratings, making it all but

impossible mathematically for analyses of inter-coder agreement to match the observed agreement (among two raters) with the expected agreement (as predicted by combinations of all possible ratings), as required to calculate Cohen's kappa. After two rounds of practice ratings of the 24 practice compositions, two coders produced percentage agreements of 76.5% for propositions, 61.7% for claims, 75.8% for data, 100% for warrants, 84.6% for oppositions, and 96.2% for responses to oppositions. For the rating of grammatical accuracy, two raters correlated at .76 (Spearman's *rho)*, which we considered adequate for our purposes.

The computer software program made mechanical the identification of strings of three words or more that appeared in each examinee's compositions and in the original source reading or listening text for that integrated writing task. Our initial trials with strings of two words or more proved to pick up too many common phrases to make the analyses meaningful, and similar trials with strings of four words or more proved to pick up too few strings from the source texts. The program also computed the total number of words in each composition (independent and integrated tasks, alike) to produce a measure of text length, then we divided that number by the number of T-units in the composition to determine the words per T-unit (as the second measure of syntactic complexity). The program further calculated the average word length in each composition by dividing the total number of characters by the total number of words in each composition, and **c**omputed a type/token ratio for each composition by dividing the total number of different words by the total number of words in a composition (excluding articles and prepositions, i.e., the following words: *a, an, the, in, on, at, of, to, for, from, off, out, into, onto, behind, above, below, over, under, along, down, up, through, across, beyond, past, before, after, since, until, about, by, with*; cf. Quirk, Greenbaum, Leech, & Svartvik, 1985). The computer program did not account for spelling errors that may have appeared in the compositions, so its calculations represent a slight understatement of the actual occurrence of specific words in the examinees' texts.

Analyses

We considered the samples of compositions to form a 3-by-3 (task type by ESL proficiency levels) factorial design, involving 36 examinees each performing 6 writing tasks (i.e., two of each task type), each administered in a randomly counterbalanced order during the field test. Our primary interest was to determine whether, for each of the nine discourse features, there were significant differences across the three task types (independent writing, reading-writing,

and listening-writing tasks) and across the three ESL proficiency levels (groups of individuals who scored either 3, 4, or 5 on the independent writing task).

We first needed to determine whether we could analyze the full set of coded data together in as few separate analyses as possible (for example, through analyses of variance) to avoid statistical error and for reasons of parsimony. To this end, we initially assessed the coded and rated data for homoscedasticity and sphericity, but we found that most of the variables did not have the same or constant variances and were not normally distributed. The data therefore violated basic assumptions for inferential statistics. But the data are appropriate for a non-parametric form of multivariate analysis of variance (NPMANOVA; Anderson, 1999, 2001; Legendre & Anderson, 1999; Legendre & Legendre, 1998, p. 19 ff; McArdle & Anderson, 2001). The NPMANOVA, originally developed for uses in ecology, does not require "normality and homogeneity of covariance matrices," and it "is relatively robust to violations of its assumptions" and to the "presence of many zeros in a data matrix" (Legendre & Anderson, 1999, p. 1). Like other nonparametric methods, the NPMANOVA is "based on measures of distance or dissimilarity between pairs of individual multivariate observations" but constructs a test statistic akin to the F-ratio in ANOVA based on "permutations of the observations to obtain a probability associated with the null hypothesis of no differences" (Anderson, 2001, p. 33).

We used a crossed design in which the fixed factors were scores on the independent writing task (as a grouping variable of ESL proficiency) crossed with three task types. Pairs of the six writing tasks were nested within one of three task types, but because we only had two samples of each writing task type, we did not have a sufficient number of specific tasks to be able to analyze systematically differences between the particular tasks within each task type (i.e., we could not assess the nested-crossed aspect of the design; cf. Neter, Wasserman, & Kutner, 1985, p. 1013). That is, our analyses focused on the crossed design. We assumed that the observations were independent and had similar distributions, satisfying the only requirement of exchangeability of the rows of the original data matrix for NPMANOVA. However, we must note that we used scores on the independent writing task (i.e., TOEFL essay) as the grouping variable because it was the only reliable, established measure of ESL writing ability available to us in the data set, though we also used this measure as one of the variables (of task type) in the analyses. We followed each multivariate test with post-hoc, pair-wise univariate tests of combinations of dependent variables, again based on permutations, so we have called these

NPANOVAs (non-parametric analysis of variance), because of their affinity to ANOVAs, in this report.

As indicated above, and described in detail in Appendices A through F, the dependent variables were:

1. text length (expressed as the total number of words per composition),
2. lexical sophistication (expressed as average word length per composition and also as a type/token ratio of the number of different lexical words over the total number of lexical words per composition),
3. syntactic complexity (expressed as the number of clauses per T-unit and as the number of words per T-unit),
4. grammatical accuracy (expressed as a holistic rating from 1 to 3),
5. argument structure (expressed as ratings from 0 to 3 of propositions, claims, data, warrants, oppositions, responses),
6. orientations to source evidence in voice (expressed as percentage of T-units per composition with unspecified voice, self as voice, specified others as voice, unspecified others as voice, or assumed community as voice),
7. the functions or content of the message in source evidence (expressed as the percentage of T-units per composition that are either declarations, quotations, paraphrases, or summaries), and
8. the extent of uses of verbatim phrases from source texts (expressed as the number of verbatim strings of three words or more from the source text).

We standardized the data to *z* scores prior to the analyses to reduce some of the disparity among the various measures. We used Bray-Curtis distances, rather than Euclidean distances, to measure the magnitudes of the dependent variables for use in the NPMANOVAs and NPANOVAs (as recommended in Anderson, 2001; Legendre & Anderson, 1999). To test for significant differences, we conducted *F* tests based on 999 restricted permutations of the standardized raw data, a method that differs from the usual *F* statistic in MANOVA.. For each analysis, we calculated the effect size using partial eta squared (η_p^2), which describes the amount

of variance accounted for in the sample (computed as the sum of squares explained divided by the sum of squares plus the error variance), as recommended by Levine and Hullett (2002) and Tabachnik and Fidell (1996, p. 54), so as to account for the effects and magnitude of additional variables on the variance for each factor in the two-way design. We interpreted these in reference to the guidelines for η_p^2 values in Cohen (1969, pp. 278-280; i.e., small effect size = 1%, medium effect size = 6%, large effect size = 14%). We should caution that we made a large number of observations on a relatively small data set, which may have introduced the possibility of statistical error. A further caution is that the NPMANOVA does not test for homoscedasticity inside subgroups, nor did we correct pair-wise tests for multiple comparisons.

We also conducted impressionistic analyses of the independent and integrated tasks. After reviewing the full set of sample compositions (and after the computer program had highlighted verbatim strings of text from the source reading or listening texts), we selected nine case-study compositions that we considered to represent ineffective, typical, and effective pieces of writing for the three task types (presented in Appendix H). For ease of reading in this report, we selected compositions from only one task within each task type, opting for the independence task to exemplify examinees' independent writing, the cinema task to represent the reading-writing task type, and the ethics/Plato task to represent the listening-writing task type. In these impressionistic analyses we tried to describe patterns that indicate how examinees used argumentation, evidence, and source texts in their written compositions. We attempted to distinguish particular ways in which this had been done, distinguishing those ways that seemed less effective from those that appeared more effective, as well as performances that seemed typical of the sample compositions as a whole.

Results

Tables 2, 3, and 4 show the means and standard deviations on each of the 6 composition tasks, and the combined means and standard deviations within each of the three task types, for each of the nine discourse indicators, plus their subcategories (i.e., within analyses of argument structure, voice, and message) for the 12 compositions at each of the Proficiency Levels 3, 4, and 5, respectively. Table 5 shows the results of the NPMANOVAs for each analysis and the corresponding effect sizes.

Text Length

For text length (i.e., total number of words written), the NPMANOVA showed a main effect, and large effect size, for task type [F (2, 207) = 154.42, p = .001, η_p^2 = .60]; a main effect, and medium effect size, for proficiency level [F (2, 207) = 28.67, p = .001, η_p^2 = .22]; and an interaction, with a small effect size, between the two factors [F (4, 207) = 5.24, p = .001, η_p^2 = .03]. The interaction indicates that there was not a consistent pattern of results for the main factors (of task type and ESL proficiency), but rather certain combinations of task types and ESL proficiency levels functioned differently for the variable of text length. This interaction effect in these NPMANOVA results appears to have arisen because the words written by examinees at ESL Proficiency Level 4 were not significantly different from those written by examinees at ESL Proficiency Level 5, although there were consistent differences in the words written between examinees at ESL Proficiency Levels 3 and 4 and between examinees at ESL Proficiency Levels 3 and 5. This trend for text length appeared across each of the three task types. Specifically, ANOVAs showed that, on the measure of text length, there were significant differences between the task types for independent writing and reading-writing (t = 11.99, p = .001), for independent writing and listening-writing (t = 14.51, p = .001), and for reading-writing and listening-writing (t = 3.90, p = .001). For proficiency level, NPANOVAs showed significant differences in text length between the groups of compositions with independent essays scored at Levels 3 and 4 (t = 3.66, p = .001) and between the groups of compositions with independent essays scored at Levels 3 and 5 (t = 4.20, p = .001), but not between the groups of compositions with independent essays scored at Levels 4 and 5 (t = .95, p = n.s.).

As shown in Tables 2, 3, and 4, the mean number of words per composition was considerably higher in the independent writing tasks (ranging from M = 274.3 for Level 3 to M = 373.0 for Level 5) than in the reading-writing tasks (ranging from M = 157.6 for Level 3 to M = 214.5 for Level 5), which were in turn higher than those in the listening-writing tasks (which range from M = 116.4 at Level 3 to M = 173.5 at Level 5). Although examinees in the field test wrote many more words in the independent essay tasks than they did in either of the integrated tasks, the mean number of words written per composition increased by proficiency level only between ESL Proficiency Levels 3 and 4 (and likewise between Levels 3 and 5), as indicated in these tables. Between ESL Proficiency Levels 4 and 5, the number of words written was not significantly different for any of the three task types.

Table 2

Means and Standard Deviations for Variables in 12 Compositions at Level 3

	Independent writing tasks						Reading-writing tasks						Listening-writing tasks					
	Independence (n = 12)		Trip plans (n = 12)		Combined (n = 24)		Politics (n = 12)		Cinema (n = 12)		Combined (n = 24)		Plato (n = 12)		Behaviorism (n = 12)		Combined (n = 24)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	SD
# of words	260.7	48.0	288.0	45.2	274.3	47.7	162.8	32.9	152.3	38.1	157.6	35.2	132.6	36.6	100.2	40.4	116.4	41.2
Word length	4.3	0.3	4.1	0.2	4.2	0.3	5.3	0.3	4.8	0.4	5.0	0.4	4.8	0.2	4.8	0.3	4.8	0.3
Type-token ratio of words	0.4	0.1	0.4	0	0.4	0	0.5	0	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1
Words per T-unit	14.6	3.8	15.6	3.7	15.1	3.7	14.7	2.7	15.6	5.4	15.2	4.2	12.1	3.5	13.9	2.5	13.0	3.1
Clauses per T-unit	1.7	0.3	1.8	0.4	1.8	0.4	1.5	0.3	1.4	0.3	1.5	0.3	1.4	0.2	1.8	0.5	1.6	0.4
# of T-units	19.3	7.2	19.3	5.4	19.3	6.2	11.2	2.0	10.5	3.7	10.8	2.9	11.9	4.4	7.3	9.6	4.4	
Verbatim phrases	0	0	0	0	0	0	11.2	6.2	9.1	5.6	10.1	5.9	3.4	2.3	3.2	1.4	3.3	1.9
Grammar	1.8	0.7	1.5	0.9	1.7	0.6	1.7	0.4	1.9	0.5	1.8	0.6	1.7	0.7	1.7	0.5	1.7	0.6
Propositions	1.9	0.5	2.2	0.4	2.0	0.5	0.8	0.9	1.1	0.8	0.9	0.8	1.1	0.9	1.3	0.7	1.2	0.8
Claims	1.6	0.5	1.8	0.7	1.8	0.6	1.5	0.7	1.3	0.5	1.4	0.6	1.6	0.5	1.2	0.4	1.4	0.5
Data	0.8	0.7	0.8	0.9	0.8	0.8	0.2	0.4	1.3	0.5	0.7	0.7	1.7	0.7	0.8	0.5	1.2	0.7

(Table continues)

Table 2 (continued)

	Independent writing tasks						Reading-writing tasks						Listening-writing tasks					
	Independence (n = 12)		Trip plans (n = 12)		Combined (n = 24)		Politics (n = 12)		Cinema (n = 12)		Combined (n = 24)		Plato (n = 12)		Behaviorism (n = 12)		Combined (n = 24)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	SD
Warrants	0.1	0.3	0.6	0.8	0.3	0.6	0.1	0.3	0	0	0	0.2	0	0	0	0	0	0
Oppositions	0.6	0.9	0.3	0.6	0.4	0.8	0	0	0	0	0	0	0	0	0	0	0	0
Responses	0.3	0.8	0	0	0.2	0.6	0	0	0	0	0	0	0	0	0	0	0	0
Unspecified voice	86.8	7.2	91.8	0.4	89.3	6.1	85.9	12.5	97.1	5.2	91.5	11.0	87.3	13.2	97.2	6.7	92.2	11.4
Self voice	12.3	7.0	8.1	3.6	10.3	5.8	0	0	2.9	5.3	1.5	4.0	0.7	2.3	0	0	0.3	1.6
Specified other voice	0.8	1.9	0	0	0.4	1.4	14.1	12.5	0	0	7.0	11.3	12.1	13.2	2.8	6.7	7.5	1.3
Unspecified other voice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Community	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Declaration	99.2	2.0	100	0	99.6	1.4	11.6	17.0	9.6	22.2	10.6	19.4	0.6	2.3	12.8	6.8	12.2	15.0
Quotation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Paraphrase	0.8	2.0	0	0	0.4	1.4	65.0	33.1	79.7	24.3	72.3	29.3	90.7	10.6	75.2	17.8	82.9	0.2
Summary	0	0	0	0	0	0	23.4	24.7	10.8	12.2	17.1	20.0	8.7	10.4	12.0	16.8	10.3	13.8

Table 3

Means and Standard Deviations for Variables in 12 Compositions at Level 4

	Independent writing tasks						Reading-writing tasks						Listening-writing tasks					
	Independence (*n* = 12)		Trip plans (*n* = 12)		Combined (*n* = 24)		Politics (*n* = 12)		Cinema (*n* = 12)		Combined (*n* = 24)		Plato (*n* = 12)		Behaviorism (*n* = 12)		Combined (*n* = 24)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# of words	336.0	69.23	334.6	65.0	335.3	65.7	212.8	40.6	190.2	51.1	201.5	46.6	182.4	29.6	159.4	22.1	170.9	28.1
Word length	4.3	0.2	4.1	0.3	4.2	0.3	5.1	0.3	4.8	0.4	5.0	0.4	4.8	0.2	4.8	0.2	4.8	0.2
Type-token ratio of words	0.4	0.1	0.4	0.1	0.4	0.1	0.5	0.1	.5	0.1	0.5	0.1	0.4	0	0.5	0.1	0.4	0.1
Words per T-unit	15.7	3.4	15.2	3.4	15.4	3.4	16.9	3.4	15.9	3.2	16.4	3.3	13.5	2.3	16.6	4.0	15.0	3.6
Clauses per T-unit	1.7	0.3	1.7	0.3	1.7	0.3	1.6	0.3	1.4	0.2	1.5	0.3	1.5	0.2	2.0	0.3	1.8	0.4
# of T-units	22.8	7.5	23.8	9.3	23.3	8.3	13.1	3.7	12.6	4.6	12.8	4.1	13.7	1.4	10.2	2.8	11.9	2.8
Verbatim phrases	0	0	0	0	0	0	12.2	6.5	5.0	2.9	8.6	6.1	7.1	3.3	6.2	2.4	6.6	2.9
Grammar	2.3	0.8	2.4	0.5	2.4	0.7	2.4	0.7	2.0	0.7	2.2	0.7	2.4	0.7	2.3	0.8	2.3	0.7
Propositions	2.0	0.4	2.0	0.4	2.0	0.4	1.8	0.8	1.2	0.8	1.5	0.8	1.3	1.2	1.2	0.9	1.3	1.0
Claims	1.8	0.6	1.8	0.6	1.8	0.6	2.1	0.7	1.7	0.7	1.9	0.7	1.7	0.7	1.7	0.7	1.7	0.6
Data	0.8	0.9	1.2	1.0	1.0	1.0	0.2	0.4	1.4	0.7	0.8	0.8	1.8	0.6	1.8	0.8	1.8	0.7

(Table continues)

Table 3 (continued)

	Independent writing tasks						Reading-writing tasks						Listening-writing tasks					
	Independence (n = 12)		Trip plans (n = 12)		Combined (n = 24)		Politics (n = 12)		Cinema (n = 12)		Combined (n = 24)		Plato (n = 12)		Behaviorism (n = 12)		Combined (n = 24)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Warrants	0.2	0.6	0.2	0.6	0.2	0.6	0	0	0	0	0	0	0	0	0.2	0.6	0.1	0.4
Oppositions	0.6	0.9	0.3	0.8	0.5	0.8	0	0	0	0	0	0	0	0	0.2	0.6	0.1	0.4
Responses	0	0	0.2	0.6	0.1	0.4	0	0	0	0	0	0	0	0	0	0	0	0
Unspecified voice	91.3	4.7	92.6	8.7	92.0	6.9	78.5	21.3	98.8	2.8	88.7	18.1	84.1	13.6	90.5	9.3	87.3	11.8
Self voice	8.3	5.0	7.3	8.7	7.8	6.9	0	0	0.5	1.7	0.3	1.2	0	0	0	0	0	0
Specified other voice	0.8	2.0	0	0	0.4	1.4	14.1	12.5	0	0	7.0	11.3	12.1	13.2	2.8	6.7	7.5	11.3
Unspecified other voice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Community	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Declaration	99.3	2.3	100	0	99.7	1.6	2.8	7.3	3.6	5.9	3.2	6.5	0	0	3.3	5.2	1.7	4.0
Quotation	0	0	0	0	0	0	2.5	8.7	7.5	2.6	1.6	6.3	0	0	0	0	0	0
Paraphrase	6.7	3.3	0	0	0.3	1.6	76.6	20.8	84.2	18.2	80.4	19.5	91.8	7.5	85.2	11.0	88.5	9.8
Summary	0	0	0	0	0	0	18.1	17.8	11.4	16.6	14.8	16.7	8.2	7.5	11.5	9.9	9.8	8.7

Table 4

Means and Standard Deviations for Variables in 12 Compositions at Level 5

	Independent writing tasks						Reading-writing tasks						Listening-writing tasks					
	Independence (*n* = 12)		Trip plans (*n* = 12)		Combined (*n* = 24)		Politics (*n* = 12)		Cinema (*n* = 12)		Combined (*n* = 24)		Plato (*n* = 12)		Behaviorism (*n* = 12)		Combined (*n* = 24)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# of words	383.4	71.0	362.6	49.0	373.0	60.6	235.3	50.6	193.7	54.4	214.5	55.6	193.3	37.0	153.8	26. 7	173.5	37.4
Word length	4.5	0.3	4.3	0.2	4.4	0.3	5.2	0.2	4.9	0.2	5.0	0.2	4.8	0.2	5.0	0.2	5.0	0.2
Type-token ratio of words	0.4	0.1	0.4	0	0.4	0.1	0.5	0	0.5	0.1	0.5	0.1	0.4	0.1	0.5	0	0.5	0.1
Words per T-unit	18.2	3.6	18.0	2.9	18.1	3.2	18.8	3.6	17.4	4.1	18.1	3.8	14.9	3.2	16.5	2.5	15.7	2.9
Clauses per T-unit	1.8	0.3	1.7	.2	1.8	0.2	1.7	0.3	1.4	0.2	1.5	0.3	1.5	0.2	2.0	0.3	1.7	0.4
# of T-units	22.1	7.3	20.7	4.7	21.4	6.0	12.8	2.6	11.3	3.2	12.0	2.9	13.5	3.7	9.6	2.7	11.5	3.7
Verbatim phrases	0	0	0	0	0	0	11.7	3.4	6.5	3.5	9.1	4.3	7.0	3.2	5.5	2.4	6.2	2.8
Grammar	2.8	0.5	2.9	0.3	2.8	0.4	2.7	0.5	2.8	0.5	2.7	0.5	2.8	0.5	2.8	0.5	2.8	0.4
Propositions	1.7	0.8	2.1	0.3	1.9	0.6	1.8	1.1	1.2	0.9	1.5	1.1	1.3	1.0	1.5	0.9	1.4	0.9
Claims	1.8	0.9	2.4	0.7	2.1	0.8	2.3	0.8	1.5	0.5	1.9	0.8	1.6	0.7	1.8	0.8	1.7	0.7
Data	1.1	0.8	1.2	1.1	1.1	1.0	0.2	0.4	1.8	0.7	1.8	0.7	2.0	0.4	1.8	0.8	1.8	0.6

(Table continues)

Table 4 (continued)

	Independent writing tasks						Reading-writing tasks						Listening-writing tasks					
	Independence (n = 12)		Trip plans (n = 12)		Combined (n = 24)		Politics (n = 12)		Cinema (n = 12)		Combined (n = 24)		Plato (n = 12)		Behaviorism (n = 12)		Combined (n = 24)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Warrants	0.2	0.6	0.1	0.5	0	0	0	0	0.2	0.6	0.1	0.4	0	0	0	0	0	0
Oppositions	0	0	0	0	0	0	0.2	0.6	0.1	0.5	0.2	0.1	0	0	0.2	0.6	0.1	0.4
Responses	0	0	0	0	0	0	0	0	0	0	0.9	0.1	.9	0.1	0.9	0.1	0.9	0
Unspecified voice	90.0	7.6	91.1	5.7	90.5	6.7	85.8	9.0	99.3	2.3	92.5	9.4	91.4	8.5	85.6	14.0	88.5	11.7
Self voice	8.6	7.5	7.6	4.2	8.2	6.0	0	0	0	0	0	0	0	0	0	0	0	0
Specified other voice	1.0	2.3	0	0	0.5	0.2	14.3	9.0	6.7	2.3	7.5	9.5	8.6	8.5	14.4	14.0	11.5	11.7
Unspecified other voice	0.3	1.2	0.6	2.0	0.5	1.6	0	0	0	0	0	0	0	0	0	0	0	0
Community	0	0	.7	1.6	0.3	1.1	0	0	0	0	0	0	0	0	0	0	0	0
Declaration	99.7	2.5	99.1	2.2	98.9	2.3	3.5	8.4	5.6	11.2	4.5	9.7	4.8	12.5	3.1	8.1	4.0	10.3
Quotation	0.5	1.7	0.3	1.2	0.4	1.4	0	0	0	0	0	0	0	0	0	0	0	0
Paraphrase	0.8	2.0	0.6	2.0	0.7	2.0	70.8	19.7	78.8	17.3	75.8	18.6	82.3	11.5	79.7	10.6	81.0	10.9
Summary	0	0	0	0	0	0	25.8	17.5	16.7	11.6	20.7	15.4	12.9	8.3	17.3	10.6	15.1	9.5

Table 5

Results of NPMANOVAs and Effect Sizes on Standardized Data

	df	*SS*	*MS*	*F*	*p*	η_p^2
Text length						
ESL proficiency	2.00	8,551.74	4,275.87	28.67	0.001	0.22
Task type	2.00	46,057.68	23,028.84	154.42	0.001	0.60
ESL X task	4.00	3,123.55	780.89	5.24	0.001	0.09
Residual	207.00	3,0869.32	149.13			
Total	215.00	88,602.29				
Lexical sophistication						
Word length						
ESL proficiency	2.00	65.17	32.58	3.34	0.04	0.03
Task type	2.00	2,549.00	1,274.50	130.68	0.001	0.56
ESL X task	4.00	17.63	4.41	0.45	0.78	0.01
Residual	207.00	2,018.86	9.75			
Total	215.00	4,650.66				
Type-token ratio						
ESL proficiency	2.00	646.42	323.21	8.83	0.00	0.08
Task type	2.00	3,715.59	1,857.79	50.75	0.00	0.33
ESL X task	4.00	74.51	18.63	0.51	0.74	0.01
Residual	207.00	7,577.53	36.61			
Total	215.00	12,014.05				
Syntactic complexity						
Words per T-unit						
ESL proficiency	2.00	3,155.55	1,577.78	12.69	0.001	0.11
Task type	2.00	1,476.04	738.02	5.94	0.005	0.05
ESL X task	4.00	272.21	68.05	0.55	0.72	0.01
Residual	207.00	25,738.10	124.34			
Total	215.00	30,641.90				

(Table continues)

Table 5 (continued)

	df	*SS*	*MS*	*F*	*p*	η_p^2
Clauses per T-unit						
ESL proficiency	2.00	177.54	88.77	1.00	0.38	0.01
Task type	2.00	2,114.90	1,057.45	11.95	0.00	0.10
ESL X task	4.00	193.72	48.43	0.55	0.70	0.01
Residual	207.00	18,320.33	88.50			
Total	215.00	20,806.49				
Grammatical accuracy						
ESL proficiency	2.00	20,512.86	10,256.43	46.12	0.00	0.31
Task type	2.00	44.03	22.02	0.10	0.95	0.00
ESL X task	4.00	466.62	116.65	0.52	0.77	0.01
Residual	207.00	46,037.50	222.40			
Total	215.00	67,061.01				
Argument structure						
Propositions						
ESL proficiency	2.00	1,972.94	986.47	0.70	0.55	0.01
Task type	2.00	28,320.94	14,160.47	10.01	0.00	0.09
ESL X task	4.00	10,749.90	2,687.47	1.90	0.09	0.04
Residual	207.00	292,861.57	1,414.79			
Total	215.00	333,905.35				
Claims						
ESL proficiency	2.00	5,299.07	2,649.54	5.94	0.00	0.05
Task type	2.00	3,771.14	1,885.57	4.23	0.01	0.04
ESL X task	4.00	639.35	159.84	0.36	0.96	0.01
Residual	207.00	9,2328.70	446.03			
Total	215.00	102,038.26				
Data						
ESL proficiency	2.00	4,796.76	2,398.38	1.18	0.31	0.01
Task type	2.00	58,025.62	2,9012.81	14.27	0.00	0.12
ESL X task	4.00	3,623.92	905.98	0.45	0.83	0.01
Residual	207.00	420,896.30	2,033.32			
Total	215.00	487,342.6				

(Table continues)

Table 5 (continued)

	df	*SS*	*MS*	*F*	*p*	η_p^2
Warrants						
ESL proficiency	2.00	787.04	393.52	0.58	0.57	0.01
Task type	2.00	14,166.67	7,083.33	10.47	0.00	0.09
ESL X task	4.00	1,959.88	489.97	0.72	0.58	0.01
Residual	207.00	14,0046.30	676.55			
Total	215.00	156,959.89				
Oppositions						
ESL proficiency	2.00	1,939.30	969.65	1.19	0.33	0.01
Task type	2.00	40,288.07	20144.03	24.70	0.00	0.19
ESL X task	4.00	2,752.06	688.01	0.84	0.51	0.02
Residual	207.00	168,842.59	815.6			
Total	215.00	213,822.02				
Orientations to source evidence: Voice						
Unspecified voice						
ESL proficiency	2.00	65.77	32.88	0.70	0.48	0.01
Task type	2.00	63.51	31.75	0.67	0.52	0.01
ESL X task	4.00	190.72	47.68	1.01	0.43	0.02
Residual	207.00	9,744.42	47.07			
Total	215.00	10,064.42				
Self as voice						
ESL proficiency	2.00	3,903.85	1,951.92	2.88	0.03	0.03
Task type	2.00	340,501.94	170,250.97	251.51	0.00	0.71
ESL X task	4.00	1,432.81	358.20	0.53	0.82	0.01
Residual	207.00	140,122.35	676.92			
Total	215.00	485,960.95				
Voice specified as other						
ESL proficiency	2.00	6,625.89	3,312.95	1.74	0.17	0.02
Task type	2.00	112,043.56	56,021.78	29.41	0.00	0.22
ESL X task	4.00	13,678.23	3,419.56	1.79	0.12	0.03
Residual	207.00	394,355.51	1,905.10			
Total	215.00	526,703.19				

(Table continues)

Table 5 (continued)

	df	*SS*	*MS*	*F*	*p*	η_p^2
Orientations to source evidence: Message						
Declarations						
ESL proficiency	2.00	0.40	0.20	1.53	0.21	0.01
Task type	2.00	26.45	13.23	101.72	0.00	0.50
ESL X task	4.00	0.21	0.05	0.41	0.78	0.01
Residual	207.00	26.92	0.13			
Total	215.00	53.98				
Paraphrases						
ESL proficiency	2.00	985.55	492.77	1.15	0.33	0.01
Task type	2.00	418,724.70	209,362.35	489.66	0.00	0.83
ESL X task	4.00	1,118.27	279.57	0.65	0.75	0.01
Residual	207.00	88,506.83	427.57			
Total	215.00	509,335.35				
Summaries						
ESL proficiency	2.00	13,661.61	6,830.81	4.54	0.01	0.04
Task type	2.00	239,834.47	119,917.24	79.72	0.00	0.44
ESL X task	4.00	11,359.50	2,839.88	1.89	0.09	0.02
Residual	207.00	311,387.13	1,504.29			
Total	215.00	576,242.71				

Lexical Sophistication

We considered two indicators of lexical sophistication. The first was average word length, for which the NPMANOVA showed a main effect, with a large effect size, for task type [F (2, 207) = 130.68, p = .001, η_p^2 = .56] and a main effect, with a small effect size, for proficiency level [F (2, 207) = 3.34, p = .04, η_p^2 = .03]. There was no interaction between the two factors [F (4, 207) = .45, p = n.s.] nor effect size for the interaction. NPANOVAs showed that, for average word length, there were significant differences between the task types for independent writing and reading-writing (t = 14.03, p = .001), for independent writing and listening-writing (t = 12.49, p = .001), and for reading-writing and listening-writing (t = 3.75, p = .001). For proficiency level, NPANOVAs showed no significant differences in average word length between the groups of compositions with independent essays scored at Levels 3 and 4

(t = .36, p = n.s.), nor between the groups of compositions with independent essays scored at Levels 3 and 5 (t = 1.30, p = n.s.), nor between the groups of compositions with independent essays scored at Levels 4 and 5 (t = 1.73, p = n.s.). Looking to Tables 2, 3, and 4, the average word lengths per composition are relatively constant across all of the task types, but they are higher for the reading-writing tasks (M = 5.0 for Levels 3, 4, and 5) than for the listening-writing tasks (ranging from M = 4.8 for Level 3 to M = 5.0 for Level 5), which are higher than for the independent writing tasks (which range from M = 4.2 at Level 3 to M = 4.4 at Level 5). In sum, examinees tended to write longer words in the two types of integrated tasks than they did in the independent writing tasks (perhaps because, as described below, some examinees tended to employ phrases directly from the source texts). This tendency was relatively constant across English proficiency levels.

The second indicator of lexical sophistication we considered was a type-token ratio of the number of different lexical words over the total number of words per composition. The NPMANOVA for this type-token ratio produced a main effect, and large effect size, for task type [F = (2, 207) 50.75, p = .001, η_p^2 = .33]; a main effect, and medium effect size, for proficiency level [F (2, 207) = 8.82, p = .001, η_p^2 = .08]; and no interaction between the two factors [F (4, 207) = .51, p = n.s.] nor any effect size for the interaction. NPANOVAs showed that, for the type-token ratio, there were significant differences between the task types for independent writing and reading-writing (t = 9.67, p = .001), for independent writing and listening-writing (t = 7.01, p = .001), and for reading-writing compared to listening-writing (t = 2.12, p = .04). For proficiency level, NPANOVAs showed significant differences in the type-token ratio between the groups of students whose independent essays were scored at Levels 3 and 4 (t= 3.35, p = .003), and between the groups whose independent essays were scored at Levels 3 and 5 (t = 2.01, p = .04), but not between the groups of compositions with independent essays scored at Levels 4 and 5 (t = 1.52, p = n.s.). As shown in Tables 2, 3, and 4, the type-token ratios tended to be higher in the two types of integrated writing tasks (M = .5) than in the independent writing tasks (M = .4), and also tended to be higher according to proficiency level. In sum, examinees tended to write more different words in the integrated tasks than they did in the independent writing tasks. This may be because they borrowed some words verbatim from the source texts in the integrated tasks, as described below, but also perhaps because examinees were asked to write about specific content, which inherently may

have involved repetition of certain words referring to that content. Across the three task types, more proficient students also tended to use more different words in their compositions. Because the time allocations differed for each writing task, it might be expected that the number of words that examinees wrote would differ as well.

Syntactic Complexity

We analyzed syntactic complexity in two ways. The first indicator was the number of words per T-unit, for which the NPMANOVA established a main effect, and medium effect size, for task type [F (2, 207) = 5.94, p = .005, η_p^2 = .05] and a main effect, and large effect size, for proficiency level [F (2, 207) = 12.69, p = .001, η_p^2 = .11], with no interaction between the two factors nor any effect size for the interaction [F(4, 207) = .55, p = n.s.]. NPANOVAs indicated that, for the number of words per T-unit, there were no differences between the task types for independent writing and reading-writing (t = .43, p = n.s), but there were statistically significant differences between the tasks for independent writing and listening-writing (t= 2.61, p = .01) and for reading-writing and listening-writing (t = 3.00, p = .002). For proficiency level, NPANOVAs showed statistically significant differences in the number of words per T-unit between the groups of compositions with independent essays scored at Levels 3 and 4 (t= 2.11, p = .03), between the groups of compositions with independent essays scored at Levels 3 and 5 (t = 4.90, p = .001), and between the groups of compositions with independent essays scored at Levels 4 and 5 (t = 2.91, p = .004). Tables 2, 3, and 4 show that the mean number of words per T-unit was lower in the independent writing tasks (ranging from M = 15.1 for Level 3 to M = 18.1 for Level 5) than in the reading-writing tasks (ranging from M = 15.2 for Level 3 to M = 18.1 for Level 5), but they were distinctly lower in the listening-writing tasks (ranging from M = 13.0 at Level 3 to M = 15.7 at Level 5). In sum, examinees wrote more words per T-units in the independent essay and reading-writing tasks than they did in the listening-writing task (though the shorter time allocated to the latter task and the demands placed on examinees' memories to recall listening material may have influenced this result). The examinees also wrote more words per T-unit, as these figures also indicate, if they were more proficient in English.

The second indicator of syntactic complexity was the number of clauses per T-unit. For this indicator, the NPMANOVA results were a main effect, and large effect size, for task type [F (2, 207) = 11.95, p = .001, η_p^2 = .10], but not a main effect for proficiency level [F (2, 207) = 1.00, p = n.s.] and not an interaction effect between the two factors [F (4, 207) = .58,

p = n.s.]. NPANOVAs revealed, for the number of clauses per T-unit, significant differences between the task types for independent writing and reading-writing (t = 5.00, p = .001) and for reading-writing and listening-writing (t = 3.47, p = .003), but not between the task types for independent writing and listening-writing (t = 1.04, p = n.s.). In sum, as shown in Tables 2, 3, and 4, the mean number of clauses per T-unit was similar across proficiency levels but varied across the task types (M = 1.5 to 1.8).

Grammatical Accuracy

For the holistic ratings of grammatical accuracy, an NPMANOVA showed no effect for task type [F (2, 207) = .10, p = .n.s.], but a main effect, and large effect size, for proficiency level [F (2, 207) = 46.12, p = .001, η_p^2 = .31], and no interaction between the two factors [F (4, 207) = .52, p = n.s.]. For proficiency level, NPANOVAs showed significant differences in grammatical accuracy between the groups of compositions with independent essays scored at Levels 3 and 4 (t= 4.8, p = .001), between the groups of compositions with independent essays scored at Levels 3 and 5 (t = 10.36, p = .001), and between the groups of compositions with independent essays scored at Levels 4 and 5 (t = 4.60, p = .001). As displayed in Tables 2, 3, and 4, the mean ratings of grammatical accuracy increased, as would be expected, by proficiency level (from M = 1.7 or 1.8 at Level 3, to M = 2.3 or 2.4 at Level 4, to M = 2.7 or 2.8 at Level 5). Within the task types, the ratings of grammatical accuracy remained relatively constant. In other words, the grammar in examinees' compositions tended to appear equally accurate in each of the three types of tasks they wrote.

Argument Structure

For argument structure, we rated separately (on scales from 0 to 3, as specified in Appendix E) the quality of the propositions, claims, data, warrants, oppositions, and responses to oppositions that appeared in the compositions. For the quality of propositions expressed, the NPMANOVA showed a main effect, and medium effect size, for task type [F (2, 207) = 10.01, p = .001, η_p^2 = .09], but no main effect for proficiency level [F (2, 207) = .70, p = n.s.], and no interaction between the two factors [F (4, 207) = 1.90, p = n.s.]. NPANOVAs indicated, for the quality of propositions, significant differences between the task types for independent writing and reading-writing (t = 4.36, p = .001) and for independent writing and listening-writing (t = 4.25, p = .001), but not for reading-writing and listening-writing (t = 0.20, p = n.s.). Tables 2,

3, and 4 show that the mean ratings of quality of propositions were relatively similar across proficiency levels (M = 1.4 to 2.0). But the independent essay tasks produced higher quality propositions, for all three proficiency groups, than the two integrated tasks did.

For the quality of claims in the compositions, the NPMANOVA indicated main effects, and moderate effect sizes, for task type [F (2, 207) = 4.23, p = .002, η_p^2 = .04] and for ESL proficiency level [F (2, 207) = 5.94, p = .005, η_p^2 = .05], but no interaction between the two factors [F (4, 207) = .39, p = n.s.]. NPANOVAs showed, for the quality of claims in arguments, statistically significant differences between the task types for independent writing and reading-writing (t = 1.96, p = .03) and for independent writing and listening-writing (t = 2.64, p = .001), but not between reading-writing and listening-writing (t = 1.0, p = n.s.). For proficiency level, NPANOVAs showed significant differences in the quality of claims between Proficiency Levels 3 and 4 (t= 2.83, p = .003) and between Levels 3 and 5 (t = 3.15, p = .001), but not between Levels 4 and 5 (t = 0.78, p = n.s.). Tables 2, 3, and 4 show that the mean ratings of quality of claims tended to be higher for the independent essays (M = 1.8, M = 1.8, and M = 2.1 for Levels 3, 4, and 5, respectively) than for the reading-writing tasks (M = 1.4, M = 1.9, and M = 1.9 for Levels 3, 4, and 5, respectively) or for the listening-writing tasks (M = 1.4, M = 1.7, and M = 1.7 for Levels 3, 4, and 5, respectively). The quality of these claims increased with ESL proficiency level.

For the quality of data examinees presented in the arguments in their compositions, an NPMANOVA showed main effects, and a large effect size, for task type [F (2, 207) = 14.27, p = .001, η_p^2 = .12], but not for proficiency level [F (2, 207) = 1.18, p = n.s.], and no interactions between the two factors [F (4, 207) = .55, p = n.s.]. NPANOVAs showed, for the quality of data in arguments, significant differences between the task types for independent writing and reading-writing (t = 5.00, p = .001) and between the reading-writing and listening-writing tasks (t = 3.50, p = .003), but no differences between the independent writing and listening-writing tasks (t = 1.04, p = n.s.). Tables 2, 3, and 4 show that the mean quality of data tended to increase by proficiency levels for the independent essays (M = 0.8, M = 1.0, and M = 1.1 for Levels 3, 4, and 5, respectively), for the reading-writing tasks (M = 0.7, M = 0.8, and M = 1.8 for Levels 3, 4, and 5, respectively), and for the listening-writing tasks (M = 1.2, M = 1.8, and M = 1.8 for Levels 3, 4, and 5, respectively). The two lower ESL proficiency groups had better quality data in the listening-writing task than in the two other tasks, but the high ESL proficiency group had better

quality data in both of the integrated tasks, compared to the independent writing task. The quality of data in the examinees' compositions seemed to vary by particular task. For example, all groups had limited data in the politics task but extensive data in the Plato task.

For the quality of warrants in the compositions, the NPMANOVA showed a main effect, and medium to large effect sizes, for task type [F (2, 207) = 10.47, p = .001, η_p^2 = .09], but no main effect for proficiency level [F (2, 207) = .58, p = .n.s.] and no interaction between the two factors [F (4, 207) = .72, p = n.s.]. NPANOVAs showed, for the quality of warrants in arguments, significant differences between the independent writing and reading-writing tasks (t = 3.23, p = 004) and between the independent writing and listening-writing tasks (t = 3.62, p = .001), but not between the reading-writing and listening-writing tasks (t = .58, p = n.s.). As displayed in Tables 2, 3, and 4, there were few warrants utilized overall in the compositions, but these tended to be more prevalent in the independent essays (M = 0.3, M = 0.2, and M = 0 for Levels 3, 4, and 5, respectively) than in the integrated tasks, where they scarcely appeared at all in either the reading-writing tasks (M = 0, M = 0, and M = 0.1 for Levels 3, 4, and 5, respectively) or the listening-writing tasks (M = 0, M = 0.1, and M = 0 for Levels 3, 4, and 5, respectively). As these numbers indicate, the integrated tasks tended to have better quality of warrants in them than the independent essays did. Curiously, the less proficient writers tended to use warrants more extensively in the independent writing tasks than the more proficient writers did.

For the quality of oppositions presented in the compositions, an NPMANOVA showed a main effect, and large effect size, for task type [F (2, 207) = 24.70, p = .001, η_p^2 = .19] but not for proficiency level [F (2, 207) = 1.19, p = .n.s.], and there was no interaction between the two factors [F (4, 207) = .84, p = n.s.] and only a small effect size for the interaction (η_p^2 = .02). NPANOVAs showed, for the quality of oppositions in arguments, significant differences between the independent writing and reading-writing tasks (t = 5.51, p = .001) and between the independent writing and listening-writing tasks (t = 4.72, p = .001), but not between the reading-writing and listening-writing tasks (t = 1.42, p = n.s.). As displayed in Tables 2, 3, and 4, the less proficient examinees (at Level 3) sometimes utilized oppositions in their independent essays (M = 0.4) but not at all in their integrated tasks. Likewise, the examinees at Level 4 used oppositions in their independent essays (M = 0.5) and occasionally in their listening-writing tasks (M = 0.1) but not at all in their reading-writing tasks. In contrast, the most proficient examinees

(at Level 5) used oppositions occasionally in their integrated tasks (M = 0.2 in the reading-writing tasks; M = 0.1 in the listening-writing tasks) but not at all in the independent essays.

For the quality of responses in the argument structures, a pattern emerged that was similar to those described above for oppositions. Statistical tests for NPANOVA and interaction effect could not be computed for the argument structures variable because of the large numbers of zero values for the response category at each ESL proficiency level. As Tables 2, 3, and 4 show, the less proficient examinees (at Level 3) occasionally used responses in their independent essays (M = 0.2) but did not use them at all in their integrated tasks. Likewise, the examinees at Level 4 sometimes used responses in their independent essays (M = 0.1) but not at all in their reading-writing tasks or listening-writing tasks. In contrast, the most proficient examinees (at Level 5) used responses occasionally in their integrated tasks (M = .9 in the reading-writing tasks; M = .9 in the listening-writing tasks) but not at all in the independent essays. Because of these erratic tendencies in uses of responses, NPANOVAs could not be run to establish differences across the particular task types.

Orientations to Source Evidence: Voice

We tallied the percentage of T-units in the compositions in which five logical options were utilized for the presentation of source information. Either (a) the source was unspecified, (b) the self was identified as the source, (c) someone else (other than the self) was identified as the source, (d) the source was indicated (as someone other than the self) but was not identified, or (e) shared, community knowledge was presented as the source.

For the percentage of T-units in which a source was not indicated or specified, the NPMANOVA did not show any main effects for task type [F (2, 207) = .67, p = n.s.] or for proficiency level [F (2, 207) = .70, p = n.s.], and there was no interaction between the two factors [F (4, 207) = 1.01, p = n.s.]. Tables 2, 3, and 4 show that the mean percentages of T-units that did not specify a source were relatively constant across tasks and across proficiency groups (M from 84% to 97% for all task types for examinees at Levels 3, 4, and 5). In sum, the vast majority of T-units that examinees wrote in their compositions in the 2002 field test did not specify any source evidence, a tendency that was consistent across task types and proficiency levels.

For the percentage of T-units in which a source was specified as the self, a non-parametric MANOVA showed a main effect, and large effect size, for task type [F (2,

207) = 251.51, p = .001, η_p^2 = .71] and a main effect, but small effect size, for proficiency level [F (2, 207) = 2.8, p = .03, η_p^2 = .03]. There was no interaction between the two factors [F (4, 207) = .55, p = n.s.]. NPANOVAs showed, for the percentages of T-units in which a source was specified as the self, statistically significant differences between the independent writing and reading-writing tasks (t= 15.94, p = .001) and between the independent writing and listening-writing tasks (t = 18.78, p = .001), but not between the reading-writing and listening-writing tasks (t = 1.35, p = n.s.). For proficiency level, NPANOVAs failed to show any statistically significant differences for the percentage of T-units with self identified as the source, nor did it show any statistically significant difference between the groups of compositions with independent essays scored at Levels 3 and 4 (t = 1.08, p = n.s.), between the groups of compositions with independent essays scored at Levels 3 and 5 (t = 1.15, p = n.s.), nor between the groups of compositions with independent essays scored at Levels 4 and 5 (t = .24, p = n.s.). Tables 2, 3, and 4 show that examinees at all proficiency levels tended to identify their selves as sources of information in about 10% of the T-units in their independent essays, but they hardly did this at all in the two integrated tasks. Examinees at Proficiency Levels 4 and 5 almost never cited themselves as a source of information in the integrated tasks, but a small number of examinees at Proficiency Level 3 did so, notably in the cinema and Plato tasks. We did not expect examinees to cite themselves as sources of information in the integrated tasks, because the task prompts did not ask for examinees' opinions or prior knowledge.

For the percentage of T-units in which a source was specified as someone other than the self, a NPMANOVA showed a main effect, and large effect size, for task type [F (2, 207) = 29.41, p = .001, η_p^2 = .22], but there was no main effect for proficiency level [F (2, 207) = 1.74, p = n.s.], and there was no interaction between the two factors [F (4, 207) = 1.79, p = n.s.] and only a small effect size for the interaction (η_p^2 = .03). NPANOVAs showed, for the percentages of T-units in which a source was specified other than the self, significant differences between the independent writing and reading-writing tasks (t= 4.89, p = .001), between the independent writing and listening-writing tasks (t = 8.19, p = .001), and between the reading-writing and listening-writing tasks (t = 2.75, p = .006). Tables 2, 3, and 4 indicate that examinees at all proficiency levels tended to identify sources of information other than themselves (i.e., from the source texts) in the integrated tasks, especially in the politics task (M = 14% of their T-units for all groups) and the Plato task (M = 9% to 12% of their T-units for all groups). But

they almost never did this in their independent essays (with the exception of a few examinees at Proficiency Levels 3 and 5 in the independence task). Examinees at Proficiency Level 5 contrasted with the other two proficiency groups by specifying sources of information from the source texts fairly often in the cinema (M = 6.7%) and the behaviorism (M = 14.4%) tasks, whereas examinees at Proficiency Levels 3 and 4 did not do this at all in the cinema task and only infrequently in the behaviorism task (M = 2.8% for both groups).

For the percentage of T-units in which a source was indicated but not specified, the NPMANOVA failed to compute because of zero values for almost all of the task types at most ESL proficiency levels. As Tables 2, 3, and 4 indicate, this option was hardly ever expressed by any examinees on any task type, except for a few examinees at Proficiency Level 5, who used it as a stylistic variant in the independent essay tasks (see example compositions discussed in Appendix H). A similar result emerged for the percentage of T-units in which a communal voice was assumed as a source of information (i.e., statistical tests for the NPANOVA main effects and the interaction effect could not be computed because of the zero values in most cells of the data matrix). As Tables 2, 3, and 4 show, only a few T-units written by a few examinees on the Trip Plans task used the option of community voice, again as a stylistic variant rather than as a consistent way of orienting their uses of source evidence.

Orientations to Source Evidence: Message

As explained in Appendix F, we tallied the percentage of T-units in the compositions in which the content of source information was presented in one of four ways, as either a (a) declaration, (b) quotation, (c) paraphrase, or (d) summary. For the percentage of T-units in which source information was presented as a declaration, an NPMANOVA produced significant main effects, and large effect sizes, for task type [F (2, 207) = 101.72, p = .001, η_p^2 = .50] but no differences for proficiency level [F (2, 207) = 1.53, p = n.s.], and there was no interaction between the two factors [F (4, 207) = .41, p = n.s.]. NPANOVAs showed, for the percentages of T-units in which source information was presented as a declaration, statistically significant differences appeared between task types for the independent writing and reading-writing tasks (t = 12.30, p = .001) and for the independent writing and listening-writing tasks (t = 16.43, p = .001), but not between the reading-writing and listening-writing tasks (t = 1.51, p = n.s.). As shown in Tables 2, 3, and 4, declarations were the predominant form of expressing the message for almost all T-units for all proficiency groups in the independent essays (M = 99% to 100%).

In contrast, in the two sets of integrated tasks, declarations only formed a small portion of the T-units for all proficiency groups (M = 3% to 13%). Examinees at Proficiency Level 3 tended to use more declarations in the integrated tasks (M = 10% to 13%) than did examinees at proficiency levels 4 and 5 (M = 0 to 6%). These usages varied, it should be observed, within each proficiency group and each task, as indicated by standard deviations that exceed the means for certain tasks in Tables 2, 3, and 4.

T-units in which source information was presented as a quotation were infrequent overall, and their appearance varied by task and by proficiency level. Because of the large number of zero values in the data matrix, the NPMANOVA would not compute results. Examinees at Proficiency Level 3 did not use quotations at all. Examinees at Proficiency Level 4 only used quotations in the reading-writing tasks (M = 3% in the politics task; M = 8% in the cinema task). Examinees at Proficiency Level 5 did not use quotations at all in the integrated tasks, but a few quotations appeared in their independent essays.

For the percentage of T-units in which source information was presented as a paraphrase, the NPMANOVA showed a main effect, and large effect size, for task type [F (2, 207) = 489.66, p = .001, η_p^2 = .83] but no effect for proficiency level [F (2, 207) = 1.15, p = n.s.], and there was no interaction between the two factors [F (4, 207) = .65, p = n.s.]. In NPANOVAs of the percentages of T-units in which source information was presented as paraphrases, there were significant differences between the independent writing and reading-writing tasks (t= 22.47, p = .001), between the independent writing and listening-writing tasks (t = 27.61, p = .001), and between the reading-writing and listening-writing tasks (t = 2.26, p = .006). Tables 2, 3, and 4 show that examinees at all proficiency levels tended to paraphrase information extensively from the source texts in the integrated tasks (M = 65% to 92% of their T-units). In contrast, paraphrases appeared only sporadically in the independent essays (M = 0% to 7%) and not at all in the Trip Plans task for examinees at Proficiency Levels 3 and 4. Of course, the independent essays did not have explicit source material in the test to paraphrase; nonetheless, in some instances examinees chose to paraphrase information from various personal sources. Examinees at Proficiency Level 4 tended to use paraphrase in about 5% to 10% more of their T-units in all of the integrated tasks than did examinees at Proficiency Levels 5 or 3.

For the percentage of T-units in which source information was presented as a summary, the NPMANOVA showed a significant main effect, and large effect size, for task type

[F (2, 207) = 79.72, p = .001, η_p^2 = .44] and a main effect, but small effect size, for proficiency level [F (2, 207) = 4.54, p = .006, η_p^2 = .04]. There was no interaction between the two factors [F (4, 207) = .52, p = n.s.] but a small effect size for the interaction (η_p^2 = .04). In NPANOVAs of the percentages of T-units that presented summaries of source information, there were significant differences between the independent writing and reading-writing tasks (t= 12.13, p = .001) and between the independent writing and listening-writing tasks (t = 12.42, p = .001), but not between the reading-writing and listening-writing tasks (t = 1.12, p = n.s.). For proficiency level, NPANOVAs showed no differences between the groups of compositions with independent essays scored at Levels 3 and 4 (t = .44, p = n.s.) nor between the groups of compositions with independent essays scored at Levels 4 and 5 (t = 1.75, p = n.s.), but there was a difference between the groups of compositions with independent essays scored at Levels 3 and 5 (t = 2.12, p = .01). Similar to the patterns described above with paraphrases, Tables 2, 3, and 4 show that summaries did not appear at all in the independent essays, but they formed a fifth to a quarter of all T-units in the integrated tasks (M = 10% to 26%). Examinees at Proficiency Level 5 tended to write about 5% to 10% more of their T-units using summaries than did examinees at Proficiency Levels 3 or 4.

Verbatim Uses of Source Texts in Integrated Tasks

We considered examinees' uses of material from the source texts, in the integrated reading-writing and listening-writing tasks, in two ways. The first approach follows that described above for our analyses of other discourse features. That is, we tallied the number of instances of verbatim uses of strings of three words or more from the source text that appeared in the examinees' compositions (as calculated by the computer program designed to identify such strings of words), then conducted t tests for the two task types (reading-writing and listening-writing) with source texts for each of the three ESL proficiency levels. We assumed that the independent essays did not have source texts, so we did not include them in this analysis.

The examinees' uses of verbatim strings of words from the source texts differed significantly between the reading-writing and listening-writing tasks at ESL Proficiency Level 3 (t = 3.92, p = .002) and at Level 5 (t = 2.36, p = .01) but not at Level 4 (t = .92, p = n.s.). As indicated in Tables 2, 3, and 4, there tended to be fewer verbatim phrases as examinees' proficiency increased for the reading-writing tasks, but more verbatim phrases as examinees' proficiency increased for the listening-writing tasks. Moreover, there were unique patterns in the

use of verbatim phrases for each of the integrated tasks, notably that examinees at proficiency Level 4 tended to use more verbatim phrases in the politics reading-writing task (M = 12.2) than in the cinema reading-writing task (M = 5.0), but about the same number of verbatim phrases in the two listening-reading tasks (M = 7.1 and 6.2 for the behaviorism and Plato tasks, respectively). In turn, examinees at Level 3 employed many verbatim phrases from both of the reading-writing tasks (M = 11.2 and 9.1 for the politics and cinema tasks, respectively) but relatively few from the listening-writing tasks (M = 3.4 and 3.2 for the behaviorism and Plato tasks, respectively).

These differences in textual borrowing behaviors may have resulted from examinees' differing degrees of comprehension of the source materials in each medium (i.e., reading vs. listening), memory factors (i.e., examinees had to recall source material they had heard during the listening task but could read it for the reading-writing task), as well as characteristics of the source texts themselves, such as their rhetorical organization or extent of factual or descriptive detail (i.e., the politics task produced more verbatim phrases for all groups than the cinema task did), or the conditions for writing (i.e., the reading-writing task was 25 minutes whereas the listening-writing task was 15 minutes). That is, the extent of verbatim phrases in these tasks appears to interact in complex ways with examinees' proficiency levels, the medium of comprehension of source materials, memory factors, and task characteristics and conditions as well. To examine these tendencies more closely, Tables 6 and 7 highlight information related to verbatim uses of source texts (as in Tables 2, 3, and 4), adding calculations of the percentages of verbatim words that appear in each composition, and comparing these by examinees' proficiency levels. As shown in Table 6, between one-fifth and one-third of the total compositions produced for the politics task involved words that were verbatim from the source text. Because the more proficient (i.e., Level 5) examinees' compositions are longer than those of the less proficient (i.e., Level 3) examinees, the percentages of these verbatim words decline by proficiency level. For the cinema task, examinees at Proficiency Level 3 employed many more words verbatim from the source reading text than did their counterparts at Proficiency Levels 4 and 5. As Table 7 shows, only about one-tenth of the words in examinees' compositions tended to be verbatim from the source texts for the listening-writing tasks. Less proficient (i.e., Level 3) examinees may not have understood the listening tasks or their vocabulary sufficiently to have been able to use verbatim phrases from those source materials.

Table 6

Verbatim Phrases From Source Texts in the Two Reading-Writing Tasks

	Politics			Cinema			Both tasks
Proficiency level (by independent writing task)	*M* verbatim phrases	*M* words verbatim from source	% of verbatim words in composition	*M* verbatim phrases	*M* words verbatim from source	% of verbatim words in composition	*M* verbatim phrases
3 (*n* = 12)	11.2	53.7	33%	9.1	55.0	37%	10.1
4 (*n* = 12)	12.2	57.7	28%	5.0	23.3	16%	8.6
5 (*n* = 12)	11.7	48.7	22%	6.5	27.6	17%	9.1

Table 7

Verbatim Phrases From Source Texts in the Two Listening-Writing Tasks

	Plato			Behaviorism			Both tasks
Proficiency level (by independent writing task)	*M* verbatim phrases	*M* words verbatim from source	% of verbatim words in composition	*M* verbatim phrases	*M* words verbatim from source	% of verbatim words in composition	*M* verbatim phrases
3 (*n* = 12)	3.4	11.6	9%	3.2	10.3	11%	3.3
4 (*n* = 12)	7.1	24.8	13%	6.2	19.7	12%	6.6
5 (*n* = 12)	6.9	23.5	12%	5.5	19.1	13%	6.2

Functional Uses of Argumentation, Evidence, and Source Texts: Nine Case Examples

To examine these discourse features holistically and impressionistically, we identified (as described in the Methods section above) nine compositions from our sample that exemplify, as cases, (a) the range of examinees' performance on the three task types and (b) compositions that are characteristically either ineffective, typical, or effective for each task type. The compositions described below and their original task instructions appear in Appendix H. We describe here examinees' functional uses of evidence, argumentation, and source materials in these compositions; we refer to and aim to exemplify the various analytic methods reported above.

Case 1. An ineffective independent essay. This composition was written by an examinee rated among the least proficient in our sample (average score of 2 on all tasks, but score of 3 on the independent essay). The essay presents a clearly stated proposition and makes several claims, but there are no data to support the claims. With the exception of "finally" in the last paragraph, there are no transitional phrases to mark the argument structure, and there is no obvious

conclusion. The writer makes numerous grammatical, spelling, and punctuation errors; uses a simple range of vocabulary; and displays little variety in sentence structure. Ironically, the essay had one of the highest ratios of clause per T-unit ratios in our sample (perhaps because of its punctuation errors). The examinee begins three sentences with "I think", presenting the self as the primary source of evidence, which conflicts in orientation with the 15 declarations in the second person ("you" or "your"), as if giving advice to the reader, and with numerous other third person declarations in the first and third paragraphs.

Case 2. A typical independent essay. This essay was written by an examinee who ranked slightly above the norm in our sample (average score of 3.8 on all tasks). The composition develops a proposition, set out in separate paragraphs, prefaced by a question that states the problem, followed by a brief summary of the main supporting argument. The rhetorical structure of the argument is distinctly marked by three enumerated body paragraphs and a concluding paragraph, though the overall effect of this marking seems formulaic. The argumentation involves three claims, and each is developed to some extent with hypothetical examples serving as data. Some concession is also made to potentially opposing points of view. Language errors are mostly minor, involving spelling, prepositions, and word choice. The vocabulary is not sophisticated, but there is a variety of well-controlled sentences. Many T-units do not specify a source of evidence (i.e., what we coded as declaration), relying frequently on second person declarations to convey the message (e.g., "you need to take responsibility").

Case 3. An effective independent essay. This essay was written by one of the most proficient writers in our sample (average score of 4.3 for all tasks). The essay presents one extended argument, developed creatively through numerous claims and supporting data, rather than in a formulaic manner (as in Case 2, immediately above). The argument structure is marked by a variety of transition phrases in four body paragraphs of varying sizes (rather than 1 main idea and 3 supporting points). The discussion adopts an academic tone, using third person throughout and avoiding both "I" and "you". Apart from some minor spelling errors, the range of vocabulary is varied and precise, and the language forms are essentially native-like, with many complex embedded syntactic structures and nominal or adverbial phrases. For instance, rather than using common phrases like "I think" to express the self as the primary source of evidence, the examinee stated, "I am of the view that …". This composition contained one of the only

T-units that we coded as using unspecified other for voice (i.e., "However, there is the argument that a young adult will never quite learn …").

Case 4. An ineffective composition in the reading-writing task. This response was written by one of the least proficient students in the sample (average score of 2 for all tasks). A proposition opens the response in an introductory paragraph alluding to supporting points. But the claims supporting the opening proposition are not developed, or even taken up, in the body paragraph, apart from references to the experience of watching movies. Instead, other points appear, giving an impression of incoherence. An attempt at an extended argument describes a personal experience that has little relevance to the main proposition that opened the composition. The sentences are structured erratically. Even though the text is relatively lengthy, there are many long, often run-on, sentences, comma splices, and inappropriate uses of punctuation. Minor spelling, word form, and morphological errors abound. The vocabulary is limited and occasionally incomprehensible. The examinee uses self as a source of evidence (e.g., "I can say that …"), which seems inappropriate for this task, given that the task instructions were to respond to the reading passage presented. Indeed, most T-units in the second and third paragraphs are declarations of personal experience, rather than references to the source material as evidence, which also seem undesirable for this task. Atypically, there are limited verbatim phrases from the source text, perhaps because the examinee did not have sufficient proficiency in English to comprehend the source reading passage fully or for its details.

Case 5. A typical composition in the reading-writing task. This response was written by an examinee who was rated slightly above the norm in our sample (average score of 3.7 on all tasks). The response is fairly brief and lacks an introduction; the main proposition is stated, instead, in the conclusion. Nonetheless, the proposition addresses both aspects of the question posed in the task prompt, although responding more specifically to the issue of economics than to the issue of changing personal experience. This provides an overall sense of coherence. But the response is not developed logically because the examinee has used many bits and pieces verbatim from the source text to compose the response. Indeed, 37% of all words in the composition are verbatim from the source reading. Minor grammatical and spelling errors are prominent, though there are few major errors that impede comprehension. The choice of words is comprehensible but limited and often borrowed from the source text, producing a sense of stilted

formality. None of the T-units have specified any source of evidence, yet all but one of the T-units have paraphrased the source text to convey their message.

Case 6. An effective composition in the reading-writing task. This response was from one of the top-ranked examinees in our sample (average score of 4.2 on all tasks). The response is notable for summarizing most of the substantive issues mentioned in the source reading, rather than paraphrasing these ideas or employing phrases verbatim from the source text. Only two phrases in the response derive explicitly from the source text, and one of these ("at the same time") is so common in ordinary usage as to hardly qualify as textual borrowing. Although the response is relatively brief (i.e., about the same length as Case 4, above) and lacking a developed paragraph format, the response appears like a series of notes that are internally coherent but not rhetorically linked together (except in the first two paragraphs). The main proposition, stated at the beginning and reiterated at the end of the composition, is supported by concise claims and relevant data. The choice of words seems original and precise, adopting an academic register. Correspondingly, most T-units are declarations in the third person, summarizing the source text but not citing it as the source of evidence. The text is written in a variety of syntactic structures, utilizing a range of verb tenses, with few minor errors.

Case 7. An ineffective composition in the listening-reading task. This response was written by one of the least proficient examinees in our sample (average score of 2 on all tasks). The rhetoric is structured, though in a formulaic manner (e.g., "First," "Secondly," "Finally"). However, the introduction fails to address the question posed in the task prompt, and the conclusion containing the examinee's proposition is difficult to perceive in the final two sentences of the final paragraph. The stated proposition and claims are weak, and the response contains mostly facts whose relevance is not clear. The examinee has used almost no phrases verbatim from the source text (i.e., only 7% of all words in the response), but this may be due to a lack of comprehension of the source passage. The language is characterized by simple sentences, clauses, and vocabulary. The first T-unit in the composition is the only example of a declaration referring explicitly to source evidence ("According to Plato"), but the idea cited does not in fact appear in the source text.

Case 8. A typical composition in the listening-reading task. This response was written by an examinee from the middle-range of our sample (average score of 3.2 on all tasks). The response has few signals of formal rhetorical structure, and it lacks a central proposition. The

examinee makes a few claims and presents numerous details, but these tend to paraphrase unanalyzed bits of data from the source text. Nonetheless, only 11% of the words in the response are verbatim from the source text (and as shown in Table 6, this is close to the average for the listening-writing tasks). There are frequent spelling, grammatical, and punctuation errors, but the clause structures show some complexity and variety, even when they are not punctuated appropriately. There is a range of vocabulary, some of it conveying an academic register. Most T-units paraphrase the message of the source text, and there are several T-units that make use of unspecified voice as evidence, assuming a shared understanding of the source text with the reader.

Case 9. An effective composition in the listening-writing task. This response was written by one of two examinees with the highest scores (average of 4.5 on all tasks) in our sample. The main proposition of the response appears in both the introductory and concluding statements, and the material in between presents an extended discussion linked by transition phrases between paragraphs. Rather than recycling chunks of data from the lecture, the response makes specific, well-connected claims related to the main proposition. Notably, the examinee has summarized, rather than paraphrased or used verbatim phrases from, the source text. There were only six phrases verbatim from the source text, representing 8% of the total words in the response. The language contains minor errors, many of which may be typographical in origin, but it also contains a variety of linguistic structures, succinct clauses, and appropriate lexical choices. The opening sentences explicitly acknowledge the source text (i.e., “In this book, The Republic, Plato explains …”).

Discussion and Implications

There were significant differences at the lexical, syntactic, rhetorical, and pragmatic levels of discourse in the written compositions produced by examinees in the independent essay and in the prototype integrated tasks (involving writing in response to source reading or listening passages) in the present field test for the new TOEFL. The results of NPMANOVAs showed main effects, mostly with large effect sizes, for task type in most of our analyses. The exceptions were the grammar rating (which we expected to be a product of ESL proficiency, not task type) and the variables with large stylistic variations among groups and thus often large numbers of infrequent or zero values that were not amenable to statistical analyses (i.e., responses in arguments, unspecified uses of voice or of communal voice in source texts, uses of quotations) or

the one instance of an interaction, for text length, between task type and ESL proficiency. The frequency of some of these discourse features was also significantly different at the three levels of English proficiency that we sampled (i.e., text length, word length, type-token ratio of different words, clauses per T-unit, uses of propositions and claims in arguments, and uses of summaries of source texts), verifying that these discourse features are integral to distinguishing these score levels attributed to examinees' writing. For many discourse features, task type and ESL proficiency level both exerted independent, consistent effects on the writing of TOEFL examinees in this field test.

To summarize the results, there were significant differences between the discourse that examinees wrote for the independent essays and the integrated reading-writing or listening-writing tasks with respect to:

- lexical sophistication (in terms of word length and different words produced),
- syntactic complexity (in terms of words per T-unit and clauses per T-unit),
- argument structure (in terms of propositions, claims, data, warrants, and oppositions),
- voice in source evidence (in terms of specifying the self or other sources as evidence), and
- message in source evidence (in terms of proportions of declarations, paraphrases, and summaries).

Examinees tended, in the integrated tasks as compared to the independent essay, to write briefer compositions, to use longer words, to use a wider range of words, to write longer clauses and more clauses, to write less argumentatively oriented texts, to indicate sources of information other than self, and to paraphrase, repeat verbatim, or summarize source information more than to make declarations based on personal knowledge. These tendencies point to substantive differences in the qualities of writing that emerged across the task types. Moreover, these findings support two major evidence claims guiding the design of these writing tasks for the new TOEFL, specifically, (a) that the independent essay task prompts examinees to produce extended written arguments and (b) that the integrated tasks prompt examinees to write about and respond to textual information (i.e., from source reading and listening passages).

The task types did not, however, have any significant bearing on the perceived grammatical accuracy of the examinees' writing or the extent to which examinees wrote statements that did not specify a voice for their sources of information. Moreover, for many of

our analyses, there were no differences in the discourse elicited for the two types of integrated tasks (reading-writing vs. listening-writing), notably for most of the features of argument structure (which did not feature much in the integrated tasks) or for examinees' distinctions between evidence arising from the self or from a source text. The two types of integrated tasks produced written discourse that shares common features; the qualities of writing they elicited from examinees appeared, in these respects, similar.

Across all three task types, examinees whose English writing was more proficient (as indicated by their scores on the two independent essays, i.e., the TOEFL essay) tended, compared to examinees whose English was less proficient (on the same measures), to write longer compositions, to use more different words, to write longer and more clauses, to demonstrate greater grammatical accuracy, to have better quality propositions and claims in their arguments, and to make more summaries of source evidence. These tendencies provide support for the scoring rubrics and levels for the independent essay task (cf. prior research on the TWE) as well as for those being developed for the integrated tasks for the new TOEFL in the sense of their distinguishing consistently these aspects of examinees' written discourse across Score Levels 3, 4, and 5. We observed a tendency in the integrated tasks for the most proficient writers to summarize or synthesize ideas coherently from source materials; for the middle-range writers to paraphrase or employ verbatim piecemeal phrases from source materials; and for the least proficient writers to not comprehend the source materials sufficiently well to be able to summarize or to paraphrase competently. This variation in comprehension of the source materials may be the reason why examinees at the middle-range of proficiency tended to use more phrases verbatim from source texts than did their counterparts who were either more or less proficient in English writing. Likewise, this interaction also points toward a distinction between the reading-writing and the listening-writing tasks: Some examinees tended, in their compositions, to use more text verbatim from the reading texts than they did from the listening tasks, perhaps because they had visual access to the printed reading passage (rather than having to rely on their memories of the listening material), they had more time for the reading task, they comprehended more of the phrases in the reading passage, or combinations of all of these conditions.

Interpreting these results, however, must acknowledge the limitations of the present research. The study involved a field test, rather than real examination conditions, so students'

motivations for writing these tasks may have differed from those experienced during actual administrations of the TOEFL. Moreover, the sample of people whose compositions we selected for the research was purposive and relatively small in number, and was not designed to represent the full range of examinees who usually take the TOEFL. A further limitation in the research is that we focused on a fairly small range of discourse indicators that we determined we could code reliably and that have precedents in prior research on written text analysis. Although the indicators we selected spanned lexical, syntactic, rhetorical, and pragmatic aspects of written texts, there are many other aspects of discourse for which the present compositions could usefully have been assessed, perhaps producing different results than those we obtained.

Implications for the New TOEFL Test and Future Research

The present findings support the inclusion of integrated reading-writing and/or listening-writing tasks as measures of English writing proficiency in the new TOEFL test. These prototype tasks allow examinees to produce written discourse that differs significantly in a variety of ways from that which they produce in the independent essay on the current TOEFL, providing an additional measure of writing ability that can be scored reliably and that interconnects English language comprehension purposefully with text production. While the independent essay allows examinees to demonstrate their abilities to form coherent written arguments, based on personal knowledge and experience, the integrated writing tasks require examinees to write compositions that summarize ideas coherently that have appeared in source texts. The integrated tasks require complex cognitive, literate, and language abilities for comprehension as well as for producing written compositions that display appropriate and meaningful uses of and orientations to source evidence, both conceptually (in terms of apprehending, synthesizing, and presenting source ideas) and textually (in terms of stylistic conventions for presenting, citing, and acknowledging sources).

The scoring rubrics used in the 2002 field test appear to have distinguished adequately between three levels of ability to write responses to source texts. But in order to understand these abilities further, and potentially to refine these scoring schemes, future research needs to study the processes of composing that examinees actually use to write from source texts. To this end, process-tracing studies (e.g., using concurrent or retrospective think-aloud verbal reports) of examinees' writing integrated tasks could usefully be undertaken, particularly to describe the ways in which examinees at each score point (a) comprehend relevant, key ideas in the source

materials, (b) decide either to use verbatim phrases, quote, paraphrase, or summarize these ideas and segments of text representing them, and (c) organize their written responses to represent these ideas in conventional rhetorical forms with appropriate stylistic devices to acknowledge source evidence. Importantly, efforts are needed to establish which uses of verbatim materials from sources are, or are not, rhetorically effective in these tasks. Such research could provide further validity evidence to define the construct of writing that is assessed in these types of integrated reading-writing and listening-writing tasks. Nonetheless, the variables we assessed in the present research have proved to be both reliable and robust, and thus they point toward aspects of examinees' written texts that may be amenable to automated scoring of compositions (cf. Shermis & Burstein, 2003). The lexical features we analyzed no doubt already feature in many automated programs for scoring writing, but the aspects of argumentation, voice in uses of source evidence, or modes of paraphrasing or summarizing source data may be useful indicators for scoring higher levels of written discourse structure.

Likewise, orientation and instructional materials should be useful, as in the *LanguEdge Courseware* (ETS, 2002, and see instructional materials at http://www.ets.org/toefl/), to facilitate examinees preparing to write integrated reading-writing and listening-writing tasks on the new TOEFL. The present study points toward important differences between examinees being able to (a) paraphrase bits of information from source texts or (b) summarize and synthesize important, relevant information in the source texts. This distinction has proved fundamental, as well, in almost all of the previous research on written summarization processes cited in the initial section of the present report. Orientation and instructional materials should be able to help examinees prepare to write integrated tasks appropriately on the test, but more importantly, should also help them to practice and use this ability in English in their academic studies, and thus produce positive washback from the test to the learning and academic performance of English language students at universities and colleges.

Whether the reading-writing and listening-writing modes of integrated tasks involve inherently different abilities is a further question for future research and the design of the new TOEFL. The present research indicated that the written discourse these two task types produce is fundamentally similar in many respects, so it may be worth considering them, for assessment purposes, as alternative varieties of similar task types or even as modes of stimuli that could be used together in complementary ways as source evidence of writing ability. Minor differences in

examinees' performance on each of the particular integrated tasks also point toward the importance of understanding better how such factors as the conditions of presentation, the characteristics of source materials, examinees' comprehension of them, and memory factors may prompt more or less effective writing and greater or lesser extents of such behaviors as verbatim uses of phrases from source texts. Characteristics to consider in designing source materials for assessment purposes may include the extent to which information is presented schematically or in rhetorically or informationally well-defined chunks, the extent of factual or descriptive detail, the range of vocabulary including relatively familiar, technical or domain-specific terms, and affective variables, like perceived interest or personal relevance. Again, process-tracing studies of examinees' thinking while composing integrated tasks could be informative for determining how these variables may influence the nature of the written texts these tasks produce as well as the indicators that might mark such discourse as more or less proficient in English.

References

Abdi, R. (2002). Interpersonal metadiscourse: An indicator of interaction and identity. *Discourse Studies, 4,* 139-145.

Anderson, M. J. (1999). Non-parametric MANOVA [Computer software]. Sydney, Australia: University of Sydney, Centre for Research on Ecological Impacts of Coastal Cities.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*, 32-46.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, *11*, 17-34.

Barton, E. L. (1993). Evidentials, argumentation, and epistemological stance. *College English*, *55*, 745- 769.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, *62,* 384-411.

Biber, D. (1988). Variation across speech and writing. Cambridge, UK: Cambridge University Press.

Biber, D. (1995). Dimensions of register variation: A corpus-linguistic comparison. Cambridge, UK: Cambridge University Press.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, *9*, 93-124.

Britt, M. & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction*, *20*, 485-522.

Brown, A., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning*, *22*, 1-14.

Carlisle, R. S. (1989). The writing of Anglo and Hispanic elementary school students in bilingual, submersion, and regular programs. *Studies in Second Language Acquisition*, *11*, 257-280.

Chenoweth, A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, *18*, 90-98.

Chipere, N., Malvern, D., Duran, P., & Richards, B. (2003, March). *Some quantifiable aspects of literacy development*. Paper presented at the annual meeting of the American Association of Applied Linguistics, Washington, DC.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, *24*, 67-87.

Connor, U. (1991). Linguistic/rhetorical measures for evaluating ESL writing. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 215-225). Norwood, NJ: Ablex.

Conrad, S., & Biber, D. (2000). Adverbial marking of stance in speech and writing. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 56-73). Oxford, UK: Oxford University Press.

Crammond, J. (1998). The uses and complexity of argument structures in student persuasive writing. *Written Communication*, *15*, 230-268.

Crismore, A., Markkanen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication*, *10*, 39-71.

Crowhurst, M., & Piche, G. (1979). Audience and mode of discourse effects on syntactic complexity in writing at two grade levels. *Research in the Teaching of English*, *13*, 101-109.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies*, *1*, 1-23.

Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing*, *8*, 73-83.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. (2004). *A teacher-verification study of prototype reading and speaking tasks for New TOEFL* (TOEFL Monograph No. MS-26, ETS RM-04-05), Princeton, NJ: ETS.

Cumming, A., Kantor, R., & Powers, D. (2001). Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph No. MS-22, ETS RM-01-22). Princeton, NJ: ETS.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while scoring ESL/EFL compositions: A descriptive model. *Modern Language Journal*, *86*, 67-96.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph No. MS-18, ETS RM-00-05). Princeton, NJ: ETS.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 72-93). Clevedon, UK: Multilingual Matters.

Cumming, A., Rebuffot, J., & Ledwell, M. (1989). Reading and summarizing challenging texts in first and second languages. *Reading and Writing*, *2*, 201-219.

Day, J. D. (1986). Teaching summarization skills: Influences of student ability level and strategy difficulty. *Cognition and Instruction*, *3*, 193-210.

Deckert, G. (1993). Perspectives on plagiarism from ESL students in Hong Kong. *Journal of Second Language Writing*, *2*, 131-148.

ETS. (2002). LanguEdge courseware: Handbook for scoring speaking and writing. Princeton, NJ: Author.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 139-155.

Faigley, L. (1979). The influence of generative rhetoric on the syntactic maturity and writing effectiveness of college freshmen. *Research in the Teaching of English*, *13*, 197-206.

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL Test of Written English* (TOEFL Research Rep. No. 64, ETS RR-98-42). Princeton, NJ: ETS.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*, 123-145.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, *41*, 337-373.

Hamp-Lyons, L., & Kroll, B. (1996). *TOEFL 2000. Writing: Composition, community, and assessment* (TOEFL Monograph Series Rep. No. 5, ETS RM-96-05). Princeton, NJ: ETS.

Haswell, R. H. (1988). Error and change in college student writing. *Written Communication*, *5*, 479-499.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, *18*, 87-107.

Howard, R. (1995). Plagiarisms, authorships and the academic death penalty. *College English*, *57*, 788-806.

Hunt, K. W. (1965). *Grammatical structures at three grade levels* (NCTE Research Rep. No. 3). Urbana, IL: The National Council of Teachers of English.

Hyland, K. (1996). Talking to the academy: Forms of hedging in science research articles. *Written Communication*, *13*, 251-281.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, *4*, 51-69.

Ivanic, R., & Camps, D. (2001). I am how I sound: Voice as self-representation in L2 writing. *Journal of Second Language Writing*, *10*, 3-33.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph No. MS-16, RM-00-03). Princeton, NJ: ETS.

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and instruction*, *7*, 161-195.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Knudson, R. (1992). Analysis of argumentative writing at two grade levels. *Journal of Educational Research*, *85*, 169-179.

Kroll, B. (1988). How college freshmen view plagiarism. *Written Communication*, *5*, 203-221.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*, 439-448.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*, 123-134.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *Modern Language Journal*, *75*, 440-448.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written composition. *Applied Linguistics*, *16*, 307-322.

Lee, W., Kantor, R., & Mollaun, P. (2002, April). *Score reliability as an essential prerequisite for validating new writing and speaking tasks for TOEFL*. Paper presented at the annual TESOL Convention, Salt Lake City, UT.

Legendre, P., & Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecology Monographs*, *69*, 1-24.

Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2nd English ed.). Amsterdam: Elsevier.

Lennon, P. (1991). Error: Some problems of definition, identification, and distinction. *Applied Linguistics*, *12*, 180-196.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*, 612-625.

McArdle, B., & Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, *82*, 290-297.

McCann, T. (1989). Student argumentative writing knowledge and ability at three grade levels. *Research in the Teaching of English*, *23*, 62-76.

McCarthy Young, K., & Leinhardt, G. (1998). Writing from primary documents: A way of knowing in history. *Written Communication*, *15*, 25-68.

Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models* (2nd ed.). Homewood, IL: Richard D. Irwin, Inc.

Pennycook, A. (1996). Borrowing others' words: Text, ownership, memory, and plagiarism. *TESOL Quarterly*, *30*, 201-230.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly*, *14*, 61-69.

Plungian, V. (2001). The place of evidentiality within the universal grammatical space. *Journal of Pragmatics*, *33*, 349-357.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, *47*, 101-143.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Reid, J. (1986). Using the Writer's Workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167-188). Alexandria, VA: TESOL.

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective.* Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.

Rifkin, B., & Roberts, F. (1995). Error gravity: A critical review of research design. *Language Learning*, *45*, 511-537.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels (TOEFL Monograph No. MS-21, ETS RM-01-03). Princeton, NJ: ETS.

Rouet, J., Favart, M., Gaonach, D., & Lacroix, N. (1996). Writing from multiple documents: Argumentation strategies in novice and expert history students. In G. Rijlaarsdam, H. Van den Bergh, & M. Couzija (Eds.), *Theories, models and methodology in writing research* (pp. 44-60). Amsterdam: University of Amsterdam Press.

Scollon, R. (1994). As a matter of fact: The changing ideology of authorship and responsibility in discourse. *World Englishes*, *13*, 33-46.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahway, NJ: Erlbaum.

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication*, *21*, 171-200.

Stansfield, C., & Ross, J. (1988). A long-term research agenda for the Test of Written English. *Language Testing*, *5*, 160-186.

Stewart, M. F., & Grobe, C. H. (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. *Research in the Teaching of English*, *13*, 207-215.

Stromso, H. I., Braten, I., & Samuelstuen, M. S. (2003). Students' strategic use of multiple sources during expository text reading: A longitudinal think-aloud study. *Cognition and Instruction*, *21*, 113-147.

Tabachnik, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Thompson, G. (1996). Voices in the text: Discourse perspectives on language reports. *Applied Linguistics*, *17*, 501-530.

Thompson, G., & Yiyun, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, *12*, 365-382.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning* (2nd ed.). New York: Macmillan.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal*, *84*, 171-184.

Wiley, J., & Voss, J. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, *91*(2), 301-311.

Wineburg, S. S. (1994). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, *83*, 73-87.

Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, *19*(4), 404-425.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Honolulu, HI: University of Hawaii at Manoa.

Yeh, S. (1998). Empowering education: Teaching argumentative writing to cultural minority middle-school students. *Research in the Teaching of English*, *33*, 49-83.

List of Appendixes

Appendix A

Average Word Length

Quality measured: Lexical sophistication

Operationalized as: $\frac{number\ of\ characters}{number\ of\ words}$

Measurement:

- Count contractions as one word whether correct or not.
- Count numbers as one word.
- Count proper nouns in English and in other languages as they are written.
- Do not count hyphenated words as single words (e.g., *well-written* = 2 words).
- Count words as they are written, even if they are incorrect (e.g., *alot* = 1 word).

Appendix B

Type-Token Ratio

Quality measured: Lexical sophistication

Operationalized as: $$\frac{\Sigma \textit{different lexical items / segment}}{\Sigma \textit{lexical items / segment}}$$

Measurement:

- Divide texts into segments roughly equaling the length of the shortest text (or if that is not feasible, find a common denominator).
- In counting lexical items, include nouns, adjectives, full verbs, and adverbs with an adjectival base, especially those with *-ly* suffix (Engber, 1995, following Quirk et al., 1985).
- In counting tokens, each lexical item counts once; so, for example, plural and singular forms *(book, books)* count as two items.
- In counting types, by contrast, inflected forms count only once; for example, *book* and *books* and *talk, talking, has talked c*ount as one item.

Appendix C

Number of Clauses per T-Unit

Quality measured: Syntactic complexity

Operationalized as: *Number of clauses/T-unit*

Measurement:

1. Count the number of T-units and clauses.
2. Divide the number of clauses by the number of T-units.
3. Use the following guidelines (adapted from Polio, 1997, Appendix C, with some modifications:
   a. A T-unit equals an independent clause with all its dependent clauses.
      Example: *My school, which I liked very much, was in Saudi Arabia.*
      1 T-unit = 1T
      1 clause / 1 clause / (continuation of 1st clause) = 2 C
      Number of Clauses / T-unit = 2.0
   b. A clause equals an **overt** subject and a **finite** verb. The following are to be counted as only one clause (and, of course, one T-unit) each:
      Example: *He left the house and drove away.*
      Example: *He wanted John to leave the house.*
   c. Run-on sentences and comma splices have as many T-units as there are independent clauses.
      Example: *My school was in Saudi Arabia, I liked it very much*
      1 T-unit / 1 T-unit = 2T
      1 clause / 1 clause = 2C
      Number of Clauses / T-unit = 1.0
      Example: *My school was in Saudi Arabia I liked it very much I will always remember it.*
      1 T-unit / 1 T-unit / 1 T-unit = 3T
      1 clause / 1 clause / 1 clause = 3C
      Number of Clauses / T-unit = 1.0

d. Follow these rules for sentence fragments:
    i. If the verb or copula is missing, count the sentence as one T-unit (and, of course, as one clause).
    Example: *My school the best in Saudi Arabia.*
    ii. If an noun phrase (NP) is standing alone, attach it to the preceding or following T-unit as appropriate.
    Example: *My school in Saudi Arabia, I liked it very much*
    (Ø T-unit) / 1 T-unit
    (Ø clause) / 1 clause
    iii. If a subordinate clause is standing alone, attach it to the preceding or following sentence but count it as a separate clause.
    Example: *I lived in Saudi Arabia. Because my father worked there.*
    Ø T-unit →1 T-unit
    1 clause / 1 clause
    Number of Clauses / T-unit = 2.0
e. When there is a grammatical subject deletion in a coordinate clause, count the entire sentence as 1 T-unit, but count each clause separately.
    Example: *First we went to our school and then went out with our friends.*
    1 T-unit
    1 clause / 1 clause
    Number of Clauses / T-unit = 2.0
f. Count *so* and *but* as coordinating conjunctions, and *so that* as a subordinating conjunction unless *so* is obviously meant.
    Example: *We go to school so that we can learn.* *(so that = so that)*
    1 T-unit
    1 clause / 1 clause
    Number of Clauses / T-unit = 2.0

Example: *We need to learn so that we go to school.(so that = so)*

1 T-unit / 1 T-unit

1 clause / 1 clause

Number of Clauses / T-unit = 1.0

g. Do not count tag-questions as separate T-units or separate clauses.

Example: *We go to school in Saudi Arabia, don't we?*

1 T-unit

1 clause

Number of Clauses / T-unit = 1.0

h. Do not count tag-statements as separate T-units or separate clauses.

Example: *We go to school in Saudi Arabia and Mary does, too.*

1 T-unit

1 clause

Number of Clauses / T-unit = 1.0

i. Count S-nodes with a deleted complementizer as a subordinate clause.

Example: *I believe that she works hard and gets good grades.*

1 T-unit

1 clause / 1 clause / 1 clause

Number of Clauses / T-unit = 3.0

j. Count direct quotes as T-units (and, of course, as separate clauses).

Example: *John said ,"I am hungry."*

1 T-unit / 1 T-unit

Count clause in parentheses as individual T-units.

Example: *I believe that she works hard (she gets good grades).*

1 T-unit / 1 T-unit

1 clause / 1 clause / 1 clause

Number of Clauses / T-unit = 1.5

Appendix D

Grammatical Accuracy

We opted for a simple, holistic rating of grammatical accuracy because numerous studies have demonstrated the difficulty (if not impossibility) of reliably classifying or evaluating the qualities of errors in the writing or speech of second-language learners (e.g., Lennon, 1991; Rifkin & Roberts, 1995), given that numerous lexical, morpho-syntactic, and semantic elements combine to produce perceptions of errors, and these vary by context and perceiver. Moreover, it is questionable if there could be a simple, direct correspondence between the frequency of errors and the quality of composition writing (Haswell, 1988). So we have chosen to devise a simple 3-point scale, akin to the one appearing in Hamp-Lyons and Henning (1991), that considers impressionistically from the viewpoint of a reader of a whole composition, the frequency, range, and gravity of errors involving grammar, punctuation, spelling, and word choice in terms of:

3 Few errors (e.g., about 1 per T-unit or less) and comprehensibility seldom obscured for reader

2 Some errors (e.g., 2 or 3 per T-unit) but comprehensible to reader,

1 Many errors (e.g., more than 3 per T-unit) often affecting comprehensibility.

Appendix E

Quality of Argument Structure

Our approach is based on Toulmin's model of argument structure (Toulmin et al., 1984), and our scoring guide is adapted from Knudson (1992) and McCann (1989).

Table E1

Scoring Guide Used to Evaluate Argument Structure Quality

Element and rating	Description
Proposition	
0	The writer does not offer a proposition.
1	The writer offers a proposition, but it does not directly address the issues or it lacks clarity.
2	The writer offers a proposition that is relevant to the issues, but it is not complete or is somewhat unclear.
3	The writer offers a proposition that is relevant to the issues, clear, and complete.
Claims	
0	The writer makes no claims.
1	The writer makes claims, but they do not directly address the proposition or they lack clarity.
2	The writer makes claims that are relevant to the proposition, but they are not complete or are somewhat unclear.
3	The write makes claims that are relevant to the proposition, clear, and complete.
Data	
0	The writer offers no data.
1	The writer offers data, but they do not directly address the claims or they lack clarity.
2	The writer offers data that are relevant to the claims, but they are not complete or are somewhat unclear.
3	The writer gives supporting data that are relevant to the claims, clear, and complete.

(Table continues)

Table E1 (continued)

Element and rating	Description
Warrants	
0	The writer does not give warrants.
1	The writer gives warrants, but they do not directly address the connection between the data and the claims, or they lack clarity.
2	The writer gives warrants that are relevant to the connection between the data and the claims, but they are not complete or are somewhat unclear.
3	The writer gives warrants that are relevant to the connection between the data and the claims, clear, and complete.
Opposition	
0	The writer does not recognize opposition.
1	The writer recognizes opposition, but it does not directly address the claims or proposition, or it lacks clarity.
2	The writer recognizes opposition that is relevant to the claims or proposition, but it is not complete or is somewhat unclear.
3	The writer recognizes opposition that is relevant to the claims or proposition, clear, and complete.
Response to opposition	
0	The writer does not offer a response to opposition.
1	The writer offers a response to opposition, but it does not directly address the opposition, or it lacks clarity.
2	The writer offers a response to opposition that is relevant to the opposition, but it is not complete or is somewhat unclear.
3	The writer offers a response to opposition that is relevant to the opposition, clear, and complete.

Appendix F

Orientations to Source Evidence Through Voice and Message

Voice (who or what is presented as the source of the language being reported). Our categories, operational definitions, and examples are adapted and modified from Thompson (1996). In coding data to establish reliability we made several changes to Thompson's categories because we sometimes had difficulty judging whether a T-unit reported on evidence from the source text or not. For example, "Plato believed that the soul has three faculties" may be a reported T-unit, whereas "Plato believes in education" may be not be. In order to avoid this ambiguity, we referred to the reported verbs presented in Thompson and Yiyun (1991). If a verb used by the examinee is included in Thompson and Yiyun's (1991) list of reported verbs, then we coded the T-unit as voicing evidence from the source text.

Table F1

Operational Definitions of Coding Categories: Voice

1	Unspecified	Voice not identified.	Nowadays children do not like mathematics. Obviously, the economy is getting worse.
2	Self	The writer is the source of information or the writer expresses ideas or thoughts that are from his/her experience.	*I promise* I won't keep you a moment longer. *I think* he was a bit shorter than you are.
3	Specified other(s)	Someone other than the writer is the source of information, which is specified.	There are two signs, *one proclaiming*: "This is the birthplace of Bill Clinton, Next President of the USA." The two cricketers deserve better, *as Graham Gooch admitted*.
4	Unspecified other(s)	The writer could have identified the source, but chose not to do so.	*It was claimed* that the platypus laid eggs. Yet now there is *a suggestion* that these purchasers will have to find a 25% down-payment.
5	Community	The writer expresses shared knowledge between the writer and the reader, so no need is seen to specify the source.	The only rescuable items were a heavy rosewood desk, eastern, and a Wellington chest whose top and side panels had split badly. *Beggars can't be choosers*. *There were lorries to the left of us and lorries to the right of us*.

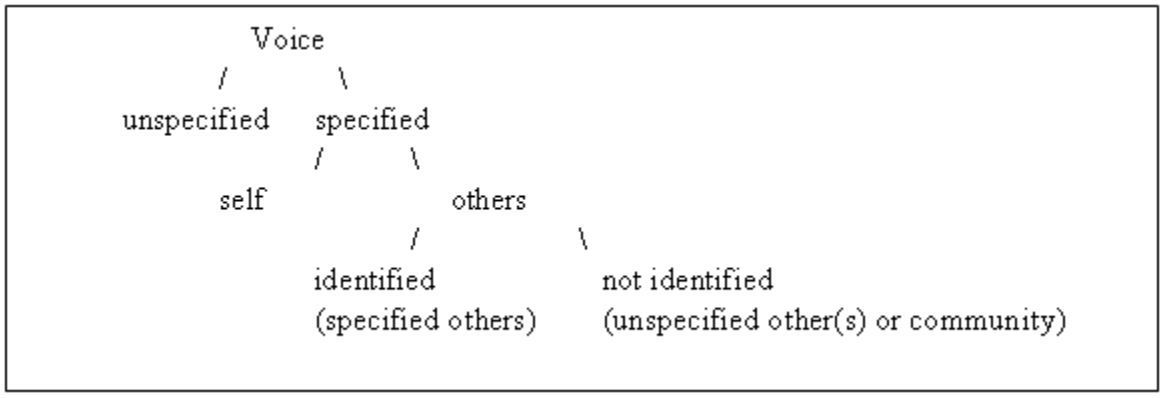

Figure F1. **Decision tree for categorizing *voice*.**

Message (the way in which the function or content of the original language is presented). In our preliminary coding to establish reliability we dropped two of Thompson's (1996) categories, *echo* and *omission*, because we did not find any examples of these functions in the sample compositions. We likewise altered Thompson's definitions of *paraphrase* and *summary*. Thompson assigned these categories only to reported speech (e.g., *paraphrase* for "Finally she asked *what I'd brought with me*" and *summary* for "Tom's boss demanded *a pledge of loyalty* from him"), concentrating on subtle differences in the form of reporting. However, for the purpose of our study, we focused more on the content of reporting; that is, how the writer used the information contained in the source passages. Hence, we decided to distinguish a T-unit as a *paraphrase* if it conveyed one idea from the source text, and as a *summary* if the T-unit made a generalization about or put together more than two ideas from the source text. This distinction between *summary* and *paraphrase* follows the definitions of previous studies of summarizing that have concluded that summarization is a more advanced and complex technique than paraphrasing (e.g., Day, 1986; Kintsch, 1990; Winograd, 1984).

Table F2

Operational Definitions of Coding Categories: Message

1	Declaration	No reference to a source text. Statement of personal opinion or fact.	I think most people are nice. Most people are nice.
2	Quotation	Verbatim quotation from source text set off by quotation marks.	"Why are you not Orthodox?" people say. Finally he lifted his chin and spoke. "I could swim when I was five."
3	Paraphrase	The writer paraphrases one idea from the source.	The workers are responsible for production in society. In an affluent society, the soldiers are needed to protect the society. The leaders are the intellectuals in the society.
4	Summary	The writer summarizes or makes a generalization about two or more ideas from the source text.	These three groups have different roles in the society.

Appendix G

Guidelines Used to Score Compositions at Levels 5, 4, and 3 for the Three Task Types

Below are the independent writing task scoring guidelines from *LanguEdge Courseware* (ETS, 2002, p. 35).

5. An essay at this level
    - effectively addresses the topic and task
    - is well-organized and well-developed, using clearly appropriate explanation, exemplification, and/or details
    - displays unity, progression, and coherence
    - displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors

4. An essay at this level
    - addresses the topic and task well, though some points many not be fully elaborated
    - is generally well-organized and well-developed, using appropriate and sufficient explanation, exemplification, and/or details
    - displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connection
    - displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure or word form or idiomatic language use that do not interfere with meaning

3. An essay at this level
    - addresses the writing topic and task using somewhat developed explanation, exemplification, and/or details
    - displays unity, progression, and coherence, though connection of ideas may be occasionally obscured
    - may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning
    - may display accurate but limited range of syntactic structures and vocabulary

Below are the reading/writing task scoring guidelines from *LanguEdge Courseware* (ETS, 2002, p. 47).

5. A response at this level has all of the following qualities:
    - principal ideas presented accurately with ample and accurately connected key supporting points/elaboration as required to fulfill the task effectively
    - organization effective in response to the task
    - sentence formation and word forms accurate and appropriate; response may have occasional minor grammatical or lexical errors
    - appropriate use of own language and language from source text

4. A response at this level has all of the following qualities:
    - principal ideas presented accurately as required by the task, though one or two key supporting points/details/elaboration may be omitted, misrepresented, or somewhat unclear, inexplicit, or inexplicitly connected
    - organization generally effective in response to the task
    - sentence formation and word choice generally accurate and appropriate; response may have noticeable minor errors and some imprecision and/or unidiomatic language use and/or imprecise connections among ideas; however, these do not obscure meaning
    - generally appropriate use of own language and language from the source text

3. A response at this level is marked by inconsistency:
    - principal ideas inconsistently presented; some are discussed accurately with key supporting points/elaboration; other support/elaboration may be absent, incorrect or unclear/obscured by weaknesses in language; **or**
    - inconsistent facility in sentence formation and word choice present (meaning may be unclear and may be occasionally obscured); **or**
    - efforts at paraphrasing may result in a number of sentence and word form errors, but meaning is not usually obscured, or there are efforts at paraphrasing, but they do not move sufficiently away from exact wordings and/or structures in the source text; **or**

- inconsistent facility in expressing connections between and among ideas (connections exist but are not effective)

Below are the listening/writing task scoring guidelines from *LanguEdge Courseware* (ETS, 2002, p. 61).

5. A response at this level
    - amply and accurately discusses the key points required by the task
    - is well-organized
    - displays accurate and appropriate sentence formation and word choice; the response may have occasional minor grammatical or lexical errors

4. A response at this level
    - accurately discusses the key points required by the task, though some key points may not be fully elaborated; the response may have occasional minor inaccuracies or distortion of information or may occasionally exhibit lack of clarity
    - is generally well-organized
    - displays generally accurate and appropriate sentence formation and word choice; the response may have noticeable minor errors and some imprecision and/or unidiomatic language use and/or imprecise connections among ideas; however these do not obscure meaning

3. A response at this level is marked by inconsistency in
    - completeness, accuracy, and/or clarity in presentation of key points; **or**
    - facility in expressing connections between and among ideas (connections exist but are not effective); **or**
    - facility in sentence formation and word choice (meaning may be unclear and may be occasionally obscured)

Appendix H

Nine Case Examples of Ineffective, Typical, and Effective Compositions for Three Task Types

Three Independent Essays

The following three compositions were written, during the 2002 field test of the new TOEFL test, to the instructions: "Read the question below. You have 30 minutes to plan, write, and revise your essay. Typically an effective response will contain a minimum of 300 words." The prompt was: "Some young adults want independence from their parents as soon as possible. Other young adults prefer to live with their families for a longer time. Which of these situations do you think is better? Use specific reasons and examples to support your opinion" (ETS, 2002 , pp. 36, 111). The compositions provided here have been reproduced exactly as they were written; hence, the spelling, typographic, spacing, and punctuation errors that appear are those of the examinees who produced the text.

Case 1. An Ineffective Independent Essay

I think that is better that young adult prefer to live with their parents for a longer time than independence from their families as soon as posible. When you are a young adult sometimes you need help to take some kind of desicions and usually you are not responsable enough to acept some situations. I think that the parents are an excellent soport in life and you do not need to move far of them because is possible tha you find soon in your way some problems.

When you life with your parents you have the opportunite to learn good things and to learn how be responsable enough when you need to take a desicion. When you life more time with them than other people you are more serious and you find less problems in your live.

Finally, I think that the parents do not want that their son or dauther live as soon as possible, then is no nesesary to move and maybe find problems and have bad time in other place where you do not know how is the real life.

Case 2. A Typical Independent Essay

Is it better to live with your parents until your 21st birthday or would it be better to move out earlier? Both opinions have their good and bad points, but generally I think moving out earlier is better. If you move out from your parents' home early, you will need to take more

responsibility for the way you spend your money, the actions you take in different circumstances and you'll also learn that making mistakes teaches and prepares you for the future.

First, you need to take responsibility of how to spend your money; you will need money for food, rent and bills. If you over spend you're in trouble. You can't spend more than you make, and this will force you to give up some things over others. You learn how to prioritize.

Second, you learn to take responsibilities of your own actions. If you are renting a place and decide to give a party to 50 of your closest friends and they brake something or make so much noise that you get a warning, you will learn to take responsibility. You will notice that your actions have consequenses and they are not always pleasant and that you can't blame anyone else but yourself. You will know what is expected of you and how you should behave to keep yourself and your neighbors happy.

Third, you will learn that life is not always easy and that making mistakes is ok, that's the way you learn things. Trial and error is the best way to learn. If you spent your money buying clothes instead of paying your phonebill, you will live without a phone until you can pay the bill. Next time, when you are faced with a choice between clothes and having a phone, you'll know better.

In conclusion, living with your parents until you are older might protect you from the world, but you will be faced with the same problems later on. It's better to learn your lessons by yourself than to have someone hand the answers to you. In the best case, a few mistakes are allowed and your parents will help you out in times of need, but even if they won't you will still learn your lesson one way or the other. It's better sooner than later.

Case 3. An Effective Independent Essay

This question is basically asking about the pros and cons of independence for the young adults as opposed to dependence on parents. I am of the view that the young adults should be be independent of their parents as soon as possible.

Young adults are basically in a position to fed for themselves. They have by law reached the age of maturity and they therefore have the basic ability to make independent choices. Whether the young adult should be allowed to be independent of their parents will depend on their level of responsibility. If the young adult exhibits maturity in their decision making process, then the person is ready to live on their own.

The main arguement is that with independence comes responsiblity. Will the young adult exhibit maturity in the choices they make regarding finances for example, dating and relationships, management of time and resources. If the young adult indicates maturity in the above (I am in no way suggesting that this list is exhaustive), then they should be allowed to be independent.

However, there is the arguement that a young adult will never quite learn responsibility and accountability unless they are let out on their own. In that regard, the young adult should be allowed as it were, to make their own mistakes and hopefully learn from them.

A good example would be a child learning how to walk. The child will have to knock a few tables, get a few bruises which will hurt for sure, but the child will eventually learn how to walk. It does not help matters if the mother of that child is consistently leading the child by the hand. It would be foolhardy to expect the child to walk on the first day. The mother will have to teach the child how to walk, guide them for a while. Then comes the practise session, when the mother has to assess as it were, the progress the child has made. The child will trip and fall to the ground, and will stumble but the child will eventually learn to walk very well with time. The speed with which the child actually picks up and learns to walk will be as a result of the mothers' teaching skills but more importantly, the childs' determination. A very determined child will learn how to walk very fast.

Applying the above example to the context of the discussion, the young adult needs proper guidance and training from the parents before being let out into the world of independence. When enough training has been given, the young adult should be given a chance to see just how much they have learnt. Initially, the parents should provide the guidance but at some point, they will need to completely let go.

Three Integrated Reading-Writing Tasks

The following three compositions were written, during the 2002 field test of the new TOEFL test in response to a passage of six paragraphs, titled “Early Cinema.” Examinees received the instructions, “You have 25 minutes to answer the question below by writing a response based on the information from the passage. Typically an effective response will be between 175 and 200 words … Explain how projectors changed the economics of showing films and the experience of watching films” (ETS, 2002, p. 147). The compositions provided here

have been reproduced exactly as they were written; hence, the spelling, typographic, spacing, and punctuation errors that appear are those of the examinees who produced the text.

Case 4. An Ineffective Composition in the Reading-Writing Task

The projectors change the economics of showing films in the way that ,people who watch films are happy , they really enjoy them and on the other hand , we have the productors of films who make too much money than before. So, i can see that , projectors have greatly change the life of everybody.Regarding , the experience of wathcing this is remarquably changed because of many aspects . Now, people can go to the movie for fun, it is like their leaving the reallity.

We can see the movie in the big screen with specials effects, it is like we communicate with the actors that's why sometimes epolple cry , or fell very sad watching the movie. For example, when i saw, the movie Titanic for the first time, i cried, i felt very sorry for all these people during many days. This the new effect of watching the movie now. The productors are in phase with people. It was something happened in the past, but it was like yesterday when i watch the movie.

In sum, i can say that projectors have changed the meaning of watch movie. This is very good aspect for the the generation of people.

Case 5. A Typical Composition in the Reading-Writing Task

The invention of projectors greatly lowered the cost of showing films. Exhibitors were able to project a handful of films to hundreds of audience at a time,charging 25 to 50 cents admission.It brought exhibitors much more profits than Kinetoscope.

With the widespread of projection,the way of watching films changed a lot.First,.motion pictures became the form of mass consumption.Audience came to see mass-produced, prerecorded materials.Second,the relationship between the image and audience was no longer private.People were sharing the pictures with even hundreds of others under the same roof.What's more,the image size was greatly enlarged from 1 or 2 inches to 6 or 9 inches.

In a word,the application of projection technology not only decreased exhibitor's cost and increased their profits,but change the way of watching films in several aspects.

Case 6. An Effective Composition in the Reading-Writing Task

Projectors significantly changed the history of film. At first, people could only watch movies by themselves, and the films were comparatively short. Projection then made it possible for large audiences to watch long motion pictures, and films became a real alternative to theater plays and ministrel shows that had atrracted mass audiences before.

Economically, this meant that a lot more people could be entertained at the same time, which makes the whole experience cheaper and probably more profitable for the owner of the movie theater.

At the same time, movie showings lost their private atmosphere. People were no longer by themselves, but they shared the expreience with many others.

Movies did not demand live actors to come to a theater and perform, everything could be prerecorded. This made it much more economical in the long run, although it also made the experience less personal.

Projectors also greatly changed the quality of movies. Images became much larger, this made the whole experiences a lot more exciting and attracted people. Increasingly, the public began to prefer movies to theater performances, a trend that we can follow until today.

Three Integrated Listening-Writing Tasks

The following three compositions were written, during the 2002 field test of the new TOEFL test in response to an audio-recorded "part of a lecture in a philosophy class" in which "the professor has been talking about ethics." Examinees received the instructions, "You have 15 minutes to answer the question below by writing a response based on the information in the lecture. Typically an effective response will be between 125 and 200 words…Plato discusses three groups of people. Using specific information from the lecture, discuss the characteristics of these three groups and their roles in society and explain the reasons why education is important for each of them" (ETS, 2002, pp. 107-108). The compositions provided here have been reproduced exactly as they were written; hence, the spelling, typographic, spacing, and punctuation errors that appear are those of the examinees who produced the text.

Case 7. An Ineffective Composition in the Listening-Writing Task

According to Plato,philosophy is very important in life.He describes the main three groups that we must have in society and how each person have to do their work.

First,we have workers who have to work hard in the purpose of reach their objectifs.In this work,their principal characteristi is desire.

Secondly,we have soldiers who are very important.They also have to work good and control their emotion.

Finally,we have leaders,the most important because they control all work and they are basis.They usually are intellectuals,they have to figure out many problems,lead workers and soldiers to do well their jobs.And the mainly point all have to do is to work together,in harmony because"union does power".Education is also a big important point in a society.

Case 8. A Typical Composition in the Listening-Writing Task

In Plate's theory, there are there groups of people in a socioty, they are workers, soldiers and leaders. The workers have a lot of desires and apitites, and their characteristcis are desire, their social roles are to produce products; soldiers are very dangerous because their social role are to protect them and their properties, so the soldiers need high spirits, so their characteristics are emotion; and the leader are these who make the decisons for the socioty, they need to use the inteligence, their roles are to lead the socioty, their characteristics are inteligent.

The workers and soldiers need to be educated to control their desire and emotiom, since without effective control, too much desire will make the workers not doing their work, and without constantly recharge the soldiers can't maitain their high spirits and then can't do their social roles well;

for the leaders, they need to be educated first , because they not only need to be inteligent but also they need to educate the workers and soldiers.

Case 9. An Effective Composition in the Listening-Writing Task

In his book, 'The Republic', Plato explains his view of a good and just person.

According to him, such qualities will only be found in a person in whom the relationship between the three parts of the soul is harmonious.

He ex, a harmonious society is th reflction of the human soul and is also made up of three different groups which all have specific role to play in the republic.

The society is therfore made up of three groups : workers, soldiers and leaders.

Workers are responisible for providing the society with basic needs. This category includes farmers and factory workers. They are intrested mainly with the basic needs of the body, such as desire.

A society also has the need to be protected and this is where soldiers are important. They have to be strong minded, high-spirited people. they carcaterise emotion.

The third category of people are the leaders whose main role it is to determine what is good for the other two categories.

However, all categories need to be taught to fully assume their role in society and this is where education plays an all important part. Self-control is essential and such moderation has to be taught.

Each cateory has to be taught to live in harmony wit the others as conflict is what mines society and this is mainly achiveable through the use of intelect. Education therefore helps control emotions or desire to create a just and good society.

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