Differences in Written Discourse in Independent and Integrated Prototype Tasks for Next Generation TOEFL

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

Robert Kantor, Educational Testing Service

Kyoko Baba, Ontario Institute for Studies in Education of the University of Toronto

Usman Erdosy, Carleton University

Keanre Eouanzoui, Niagara University

Mark James, University of Puerto Rico

Version of January 31, 2005

Contact for authors:

Alister Cumming
acumming@oise.utoronto.ca
Modern Language Centre
Ontario Institute for Studies in Education
University of Toronto
252 Bloor Street West
Toronto, Ontario
Canada M5S 1V6

Acknowledgements

This research was funded by the TOEFL research program at Educational Testing Service. The opinions expressed are those of the authors rather than the policies of the TOEFL or of Educational Testing Service. An extended version of this article will appear as a TOEFL Research Report. We thank Dan Eignor for helpful suggestions on our reporting of statistics as well as Don Banh 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.
Differences in Written Discourse in Independent and Integrated Prototype Tasks for Next Generation TOEFL

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 Next Generation 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 non-parametric MANOVAs followed a 3 (task type: TOEFL Essay, writing in response to a reading passage, writing in response to a listening passage) by 3 (English proficiency level: score levels 3, 4, and 5 on the TOEFL Essay) within-subjects factorial design. The discourse produced for the integrated writing tasks differed significantly from the discourse produced in the independent essay for the variables of: lexical complexity (text length, word length, ratio of different words to total words written), syntactic complexity (number of words per T-unit, number clauses per T-unit), rhetoric (quality of propositions, claims, data, warrants, and oppositions in argument structure), and pragmatics (orientations to source evidence in respect to self or others and to phrasing the message as either declarations, paraphrases, or summaries). Across the three English proficiency levels, significant differences appeared for the variables of grammatical accuracy as well as all indicators of lexical complexity (text length, word length, ratio of different words to total words written), one indicator of syntactic complexity (words per T-unit), one rhetorical aspect (quality of claims in argument structure), and two pragmatic aspects (expression of self as voice, messages phrased as summaries).
A fundamental issue in the assessment of writing abilities through multiple tasks is establishing that the writing elicited in a test is, on the one hand, substantially different from one task type to another, yet on the other hand, is also consistent across writing tasks. This issue came to the fore in the present study, which we conducted as research preliminary to the design of new writing tasks for the TOEFL® (Test of English as a Foreign Language) test, known as the Next Generation TOEFL® and TOEFL®/iBT (internet-based test). In particular, we focused on task types that involve examinees writing to integrate information from reading and listening stimulus materials (i.e., integrated tasks) in addition to composing the essay task that now features in the TOEFL as the TOEFL Essay® (i.e., independent task, formerly the Test of Written English, which does not require references to source materials in the test materials). Numerous arguments and sources of evidence have suggested that adding such integrated 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, Kantor, Powers, Santos & Taylor, 2000; Cumming, Kantor & Powers, 2001, 2002; Cumming, Grant, Mulcahy-Ernt & Powers, 2004; Hamp-Lyons & Kroll, 1997; Lee, Kantor & Mollaun, 2002; Rosenfeld, Leung & Oltman, 2001).

As the major test of English language proficiency for decisions about admissions to English-medium universities and colleges in North America, the TOEFL exerts a consequential, high-stakes role for students from non-English countries seeking to pursue academic studies in the United States and Canada as well as in certain other parts of the world—as do other tests with similar purposes, such as the Integrated English Language Testing System (IELTS) or the Michigan English Language Assessment Battery (MELAB). The development of the Test of Written English (TWE) in the 1980s—in response to appeals by educators—added to the TOEFL an optional, direct assessment of writing abilities (Stansfield & Ross, 1988). Recently, the TWE has become, as the TOEFL Essay, an integral element of the TOEFL (complementing other components of the test that assess reading and listening comprehension as well as vocabulary and
grammatical knowledge). The TWE or TOEFL essay has the limitations of many other holistically-scored direct assessments of writing. The single sample of argumentative-type writing elicited for the TWE or TOEFL Essay, and the resulting single score on a six-point scale, is a limited representation of an individual’s writing abilities (Cumming, 1997; Huot, 1990; Purves, 1992; Raimes, 1990). The TWE or TOELF Essay’s genre of timed, argumentative writing is but one form of written text production among the many that are actually required for successful performance in academic studies in English, and its rhetorically formulaic requirements may be prone to coaching or a reduction in the expectations for learning and teaching (i.e., negative washback) in courses for students preparing for university admissions (Alderson & Hamp-Lyons, 1996; Bailey, 1999; Hamp-Lyons & Kroll, 1997; Raimes, 1990).

Aiming to address these issues in the context of developing a new version of the TOEFL, Cumming, Kantor, Powers, Santos and Taylor (2000) proposed a series of initial, formative analyses as a basis to design new writing tasks for the test. Subsequently, Cumming, Kantor and Powers (2002) established that experienced raters of written composition from diverse backgrounds used fundamentally similar, but slightly more complex, decision-making behaviors to evaluate students’ performance on the new content-responsive, integrated tasks as they did when scoring TOEFL essays. Later, Cumming, Grant, Mulcahy-Ernt and Powers (2004) sought (and received relatively positive) formative evaluations of the new writing (and also speaking) tasks from experienced ESL instructors at several universities in regards to the authenticity and representativeness of their students’ performance on the prototype tasks. These initial studies were performed in conjunction with many other studies undertaken at Educational Testing Service (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. These analyses culminated in the publication of the LanguEdge Courseware (Educational Testing Service, 2002), which describes and exemplifies these new task types, and provides practice test tasks and instructional materials, to help prepare examinees and English language educators for Next Generation TOEFL when it becomes operational in the near future.
Prior to the present research, little systematic evidence has been 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, and how, the writing of examinees differs across these task types is vital to validating these new writing tasks, particularly in justifying their value on the test, for the design and interpretation of scoring schemes for the writing component of the test, and for the preparation of orientation and training materials for Next Generation 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 Next Generation 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, Kantor, Powers, Santos, and Taylor (2000).

Methods

The present 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 colleagues had previously pursued in analyses of the prototype writing tasks, for example, based on the verbal reports of experienced raters performing holistic evaluations of examinees’ essays
in Cumming, Kantor, and Powers (2001, 2002) or of experienced teachers judging the writing of their students on prototype tasks as in Cumming, Grant, Mulcahy-Ernt, and Powers (2004). Those previous studies had followed the premises of the reader-writer model for validating new task types for TOEFL—proposed in Cumming, Kantor, Powers, Santos, and Taylor (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, Kantor, Powers, Santos, and Taylor (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 English as a Second Language (ESL) composition assessment, aiming to select a suitable range of discourse indicators for our project. We discussed possible indicators critically among ourselves and then tried them out on sample compositions (described below) to determine that coding them was feasible and meaningful and could reliably be done on the present data. We wanted to select indicators that:

(a) 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, Kantor, Powers, Santos, and Taylor (2000) or Cumming (2001);

(b) 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;
(c) 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 which have clear theoretical justifications and operational definitions; and

(d) 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 Next Generation TOEFL, and thus also be relevant to the writing produced by adult ESL examinees who take the TOEFL.

We decided on nine indicators to apply in our discourse analyses, building on precedents established in prior studies of ESL writing and writing assessment as well as our own discussions and preliminary coding analyses on sample compositions from the field test:

1. *Text length*, operationalized as the total number of words written in a composition within the time allocated for each task (Chenoweth & Hayes, 2001; Grant & Ginther, 2000; Homburg, 1984; Perkins, 1980; Polio, 1997; Wolfe-Quintero, Inagaki, & Kim, 1998). We adopted the operational definitions appearing in Polio (1997).

2. *Lexical sophistication*, analyzed in two ways:

   2.1 *average word length* (Biber, 1988; Chipere, Malvern, Duran & Richards, 2003; Engber, 1995; Frase, Faletti, Ginther & Grant, 1999; Grant & Ginther, 2000). Our operational definition followed Engber (1995) as the number of characters divided by the number of words per composition.

   2.2 *type/token ratio* of the number of different lexical words over the total number of words per composition (Cumming & Mellow, 1996; Chipere, Malvern, Duran & Richards, 2003; Engber, 1995; Grant & Ginther, 2000; Laufer & Nation, 1995). Our operational definition followed Engber (1995).
3. Syntactic complexity, analyzed in two ways, following the operational definitions in Polio (1997) (i.e., a T-unit is an independent clause with all of its dependent clauses):

   3.1 number of clauses per T-unit (Bardovi-Harlig & Boffman, 1989; Faigley, 1979; Homburg, 1984; Ishikawa, 1995; Perkins, 1980; Polio, 1997; Wolfe-Quintero, Inagaki, & Kim, 1998);

   3.2 words per T-unit (Chipere, Malvern, Duran & Richards, 2003; Polio, 1997).

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). We operationalized this as a rating scale following the manner of Hamp-Lyons and Henning (1991), and recognizing the problems associated with evaluations or classifications or frequency counts of errors in learners’ texts (Haswell, 1988; Lennon, 1991; Rifkin & Roberts, 1995).

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 for each composition, following the operational definitions and rating scale developed by Knudson (1992) and McCann (1989), based on Toulmin, (1958) and Toulmin, Rieke, and Janik (1984) (cf. Connor, 1990; Crammond, 1998).

6. Orientations to source evidence expressed in each T-unit, coding for Presentation of Voice (as either references that are unspecified or are specified in 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 were adapted from research on reported speech by Thompson (1996) and Plungian, (2001) and on summary writing by Day (1986), Kintsch (1990), and Winograd (1984). We expected the independent essays to rely primarily on source evidence from students’ own experiences, presented in the form of declarations without sources specified, whereas we expected the integrated tasks to produce
evidence—through quotations, paraphrases, or summaries—in examinees’ compositions primarily from the source reading or listening material.

7. Functional uses of phrases from source (a) reading and (b) listening prompts that appear as verbatim strings of words in ESL compositions. We hired a computer programmer to prepare a software program that identified all strings of 3 words or more that appeared in examinees’ compositions and that also appeared in the source reading or listening materials for these integrated tasks in the field test. We used the computer program to tally these instances of verbatim phrases from the source texts then reviewed all of our sample compositions to identify, and attempt to interpret impressionistically, how examinees used the verbatim phrases from the source texts in their compositions. These analyses focused only on the reading-writing and listening-writing tasks because there were no source texts for the independent writing task.

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 types of writing that examinees produced, rather than just selected categories of words or phrases. For example, studies such as Grant and Ginther (2000), among many others, have coded 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 based our selection of discourse features on the evidence claims guiding the design of the prototype tasks for Next Generation TOEFL (i.e., as described in the LanguEdge materials as well as Cumming, Kantor, Powers, Santos & Taylor, 2000). These evidence claims focus on higher-order rhetorical functions, such as argument structure or message functions, which subsumed other forms of attitudinal expression that we might have analyzed (e.g., aspects of expressing identity are contained in our analyses of voice).
Second, our initial reviews of the composition data determined that there were too few instances of explicit markers of concepts or functions like individual identity or hedging, or these were so ambiguously expressed in the compositions, that they could not be interpreted reliably across the range of compositions in our sample. A reason for this may be, as demonstrated in 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). In the context of writing for an exam like Next Generation TOEFL, however, virtually all of these contextual factors have been eliminated in the interest of making the content and conditions of the test fair to examinees with diverse backgrounds internationally (cf. Cumming, 2002). Our selection of discourse indicators 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. Third, we wanted to take a neutral perspective on examinees' verbatim uses of phrases from the source reading or listening texts, without prejudging how or why they may have used such phrases in their writing. As numerous authors have demonstrated, plagiarism is problematic 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, academic, or institutional setting (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; McCarthy Young & Leinhardt, 1998).

Selection of Sample Compositions
We identified for analysis a purposive sample of 216 compositions written for 6 tasks in Field Test 3.1 of Next Generation TOEFL 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 had collapsed the score levels of 1 and 2 normally used for the TOEFL Essay together into a score level of 1, because (a) these score levels could not be readily distinguished among the population writing the field test and (b) to make a 5 point scale equivalent to the 5 point scale used for rating the integrated tasks in the field test (as shown in Appendix C). 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 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 worked on scoring 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 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 of these writing samples. Only 12 compositions at level 5 met this criterion, so we set the number 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 TOEFL Essay 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) and are below the proficiency levels usually considered for university admissions decisions. So we did not sample from compositions scored at these levels.

We next identified the compositions (which had already been scored by staff at ETS) that these 36 people wrote for two reading-writing and two listening-writing tasks in the 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). Examinees had been given 30 minutes to write each of the independent essays, 25 minutes to write each of the integrated reading-writing tasks, and 15 minutes to write each of the integrated listening-writing tasks. The mean scores given to the independent essays (by raters at ETS) for all 300 students in the field test tended to be about one point higher (Independence, $M = 3.22$; Plan a Trip, $M = 3.17$) than the scores given to either the reading-writing tasks (Early Cinema, $M = 2.26$; Nineteenth-century Politics, $M = 2.05$) or the listening-writing tasks (Behaviorism, $M = 2.26$; Plato, $M = 2.54$). This numerical difference appears to be due to the differences in the scaled scores described above.

The prompts used in Field Test 3.1 for these tasks have now been published in the *LanguEdge Courseware* (Educational Testing Service, 2002), so we refer readers there for details of the task instructions, rating scales, and source texts. But for ease of reference, we reproduce the scoring guidelines for the three task types in Appendix B. The instructions and prompts for each task were:

- For the independent essays, examinees were instructed to: “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 for the essay on Independence 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.” The prompt for the essay on Plan a Trip was: “Some people like to plan their trips or vacations carefully. Other people do not make detailed plans for vacations. Which approach do you prefer? Use specific reasons and examples to support your opinion.” (LanguEdge Courseware: Handbook for Scoring Speaking and Writing, pp. 36, 111, 148)

- For the integrated reading-writing task on Early Cinema, examinees received a reading passage of six paragraphs, titled “Early Cinema,” and 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.” For the integrated reading-writing task on Nineteenth-century Politics, examinees received the same instructions (as for the Early Cinema task) in reference to a reading passage of six paragraphs, titled “Nineteenth-century Politics” but the question, “Explain the different viewpoints held by nineteenth-century Whigs and Democrats on the economy and the role of government and explain why different people supported each of the two parties.” (LanguEdge Courseware: Handbook for Scoring Speaking and Writing, pp. 109, 110, 147).

- For the integrated listening-writing task on Plato, examinees heard an audio-recorded “part of a lecture in a philosophy class” in which “the professor has been talking about ethics,” and 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.” For the integrated listening-writing task on Behaviorism, examinees heard an audio-recorded “part of a psychology class” in which “the professor is discussing behaviorism,” and received the same instructions (as for the Plato task) but with
the question, “Describe laryngeal habits and ideomotor action and explain why these behaviors were of interest to psychologists such as Watson and James” (LanguEdge Courseware: Handbook for Scoring Speaking and Writing, pp. 107-108, 145, 146).

Coding, Rating, and Tallying Data

To establish inter-coder reliability prior to 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 among the research team 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 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 B 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 s/he was most familiar, experienced, and 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. We found that judging the quality of argument structures and orientations to source evidence involved more interpretations than did the other indicators, and they were difficult to reach a high inter-coder agreement on (as described below), so we made 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.

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, the coding of voice and message in each T-unit was relatively consistent, producing percentage agreements of 94% for voice categories and of 84% for the message categories (or Cohen’s kappa of .75 and .72, respectively). Rating the sub-categories 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 and, second, some of the score points on the scales (adopted from McCann, 1989 and Knudson, 1992) 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 computed 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). 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 (task type) by 3 (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 counter-balanced order during the field test. Our primary interest was to determine whether, for each of the 9 discourse features, there were significant differences across the 3 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). The NPMANOVA, originally developed for uses in ecology, does not require “normality and homogeneity of covariance matrices” and “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 non-parametric 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-cROSed aspect of the design). 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.

As indicated above, 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 in voice, or assumed community in 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 3 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 which we should observe for NPMANOVA differs from the usual $F$ statistic in MANOVA. For each analysis we calculated the effect size using partial eta squared ($\eta_p^2$), 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), 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 eta squared 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.

**Results**

Table 1 shows the results of the NPMANOVAs and the corresponding effect sizes 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 English proficiency levels 3, 4, and 5. For the sake of brevity we describe only those results which proved to be statistically significant. Full descriptive statistics from all of these analyses are forthcoming in the TOEFL Monograph report on this study (Cumming, Kantor, Baba, Erdosy, Eouanzoui, & James, forthcoming). To exemplify the qualities of writing that emerged for each task type, and aspects of the coding categories we used, at the end of this section we describe characteristics of compositions that were typically ineffective and effective for each task type.

| Table 1 |

**Results of NPMANOVAs and Effect Sizes on Standardized Data**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>$\eta_p^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text Length</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>8551.74</td>
<td>4275.87</td>
<td>28.67</td>
<td>0.001</td>
<td>0.22</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>46057.68</td>
<td>23028.84</td>
<td>154.42</td>
<td>0.001</td>
<td>0.60</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>3123.55</td>
<td>780.89</td>
<td>5.24</td>
<td>0.001</td>
<td>0.09</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>30869.32</td>
<td>149.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>88602.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lexical Sophistication</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Word Length</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>65.17</td>
<td>32.58</td>
<td>3.34</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>2549.00</td>
<td>1274.50</td>
<td>130.68</td>
<td>0.001</td>
<td>0.56</td>
</tr>
</tbody>
</table>
### ESL X Task

<table>
<thead>
<tr>
<th></th>
<th>4.00</th>
<th>17.63</th>
<th>4.41</th>
<th>0.45</th>
<th>0.78</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>2018.86</td>
<td>9.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>4650.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type-token Ratio</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>646.42</td>
<td>323.21</td>
<td>8.83</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>3715.59</td>
<td>1857.79</td>
<td>50.75</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>74.51</td>
<td>18.63</td>
<td>0.51</td>
<td>0.74</td>
<td>0.01</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>7577.53</td>
<td>36.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>12014.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Syntactic Complexity

#### Words per T-unit

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>3155.55</td>
<td>1577.78</td>
<td>12.69</td>
<td>0.001</td>
<td>0.11</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>1476.04</td>
<td>738.02</td>
<td>5.94</td>
<td>0.005</td>
<td>0.05</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>272.21</td>
<td>68.05</td>
<td>0.55</td>
<td>0.72</td>
<td>0.01</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>25738.10</td>
<td>124.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>30641.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Clauses per T-unit

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>177.54</td>
<td>88.77</td>
<td>1.00</td>
<td>0.38</td>
<td>0.01</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>2114.90</td>
<td>1057.45</td>
<td>11.95</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>193.72</td>
<td>48.43</td>
<td>0.55</td>
<td>0.70</td>
<td>0.01</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>18320.33</td>
<td>88.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>20806.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Grammatical Accuracy

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>20512.86</td>
<td>10256.43</td>
<td>46.12</td>
<td>0.00</td>
<td>0.31</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>44.03</td>
<td>22.02</td>
<td>0.10</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>466.62</td>
<td>116.65</td>
<td>0.52</td>
<td>0.77</td>
<td>0.01</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>46037.50</td>
<td>222.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>67061.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Argument Structure

#### Propositions

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>1972.94</td>
<td>986.47</td>
<td>0.70</td>
<td>0.55</td>
<td>0.01</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>28320.94</td>
<td>14160.47</td>
<td>10.01</td>
<td>0.00</td>
<td>0.09</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>10749.90</td>
<td>2687.47</td>
<td>1.90</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>292861.57</td>
<td>1414.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>333905.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Claims

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>5299.07</td>
<td>2649.54</td>
<td>5.94</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>3771.14</td>
<td>1885.57</td>
<td>4.23</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>639.35</td>
<td>159.84</td>
<td>0.36</td>
<td>0.96</td>
<td>0.01</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>92328.70</td>
<td>446.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>102038.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Data

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL proficiency</td>
<td>2.00</td>
<td>4796.76</td>
<td>2398.38</td>
<td>1.18</td>
<td>0.31</td>
<td>0.01</td>
</tr>
<tr>
<td>Task type</td>
<td>2.00</td>
<td>58025.62</td>
<td>29012.81</td>
<td>14.27</td>
<td>0.00</td>
<td>0.12</td>
</tr>
<tr>
<td>ESL X Task</td>
<td>4.00</td>
<td>3623.92</td>
<td>905.98</td>
<td>0.45</td>
<td>0.83</td>
<td>0.01</td>
</tr>
<tr>
<td>Residual</td>
<td>207.00</td>
<td>420896.30</td>
<td>2033.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>215.00</td>
<td>487342.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warrants</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>787.04</td>
<td>393.52</td>
<td>0.58</td>
<td>0.57</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td>Task type</td>
<td>2.00</td>
<td>14166.67</td>
<td>7083.33</td>
<td>10.47</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ESL X Task</td>
<td>4.00</td>
<td>1959.88</td>
<td>489.97</td>
<td>0.72</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>207.00</td>
<td>140046.30</td>
<td>676.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.00</td>
<td>156959.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oppositions</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>1939.30</td>
<td>969.65</td>
<td>1.19</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Task type</td>
<td>2.00</td>
<td>40288.07</td>
<td>20144.03</td>
<td>24.70</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ESL X Task</td>
<td>4.00</td>
<td>2752.06</td>
<td>688.01</td>
<td>0.84</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>207.00</td>
<td>168842.59</td>
<td>815.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.00</td>
<td>213822.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientations to Source Evidence: Voice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unspecified Voice</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>65.77</td>
<td>32.88</td>
<td>0.70</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>Task type</td>
<td>2.00</td>
<td>63.51</td>
<td>31.75</td>
<td>0.67</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>ESL X Task</td>
<td>4.00</td>
<td>190.72</td>
<td>47.68</td>
<td>1.01</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>207.00</td>
<td>9744.42</td>
<td>47.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.00</td>
<td>10064.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self as Voice</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>3903.85</td>
<td>1951.92</td>
<td>2.88</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Task type</td>
<td>2.00</td>
<td>340510.94</td>
<td>170250.97</td>
<td>251.51</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ESL X Task</td>
<td>4.00</td>
<td>1432.81</td>
<td>358.20</td>
<td>0.53</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>207.00</td>
<td>140122.35</td>
<td>676.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.00</td>
<td>485960.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voice Specified as Other</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>6625.89</td>
<td>3312.95</td>
<td>1.74</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>Task type</td>
<td>2.00</td>
<td>112043.56</td>
<td>56021.78</td>
<td>29.41</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ESL X Task</td>
<td>4.00</td>
<td>13678.23</td>
<td>3419.56</td>
<td>1.79</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>207.00</td>
<td>394355.51</td>
<td>1905.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.00</td>
<td>526703.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientations to Source Evidence: Message</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Declarations</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>0.40</td>
<td>0.20</td>
<td>1.53</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Task type</td>
<td>2.00</td>
<td>26.45</td>
<td>13.23</td>
<td>101.72</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ESL X Task</td>
<td>4.00</td>
<td>0.21</td>
<td>0.05</td>
<td>0.41</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>207.00</td>
<td>26.92</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.00</td>
<td>53.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paraphrases</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>985.55</td>
<td>492.77</td>
<td>1.15</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Task type</td>
<td>2.00</td>
<td>418724.70</td>
<td>209362.35</td>
<td>489.66</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ESL X Task</td>
<td>4.00</td>
<td>1118.27</td>
<td>279.57</td>
<td>0.65</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>207.00</td>
<td>88506.83</td>
<td>427.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.00</td>
<td>509335.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summaries</td>
<td>ESL proficiency</td>
<td>2.00</td>
<td>13661.61</td>
<td>6830.81</td>
<td>4.54</td>
<td>0.01</td>
</tr>
</tbody>
</table>
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 \( (F(2, 207) = 154.42, p = .001, \eta_p^2 = .60) \), a main effect, and medium effect size, for proficiency level \( (F(2, 207) = 28.67, p = .001, \eta_p^2 = .22) \), and an interaction, with a small effect size, between the two factors \( (F(4, 207) = 5.24, p = .001, \eta_p^2 = .03) \). 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 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 \( (t = 4.20, p = .001) \).

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).

**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 ($F(2, 207) = 130.68, p = .001, \eta^2_p = .56$) and a main effect, with a small effect size, for proficiency level ($F(2, 207) = 3.34, p = .04, \eta^2_p = .03$). 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 ($t = 3.75, p = .001$). The average word lengths per composition were relatively constant across all of the task types, but they were higher for the reading-writing tasks ($M = 5.0$ for levels 3, 4, and 5) than the listening-writing tasks (ranging from $M = 4.8$ for level 3 to $M = 5.0$ for level 5), which were 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 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 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 ($F = (2, 207) 50.75, p = .001, \eta^2_p = .33$) and a main effect, and medium effect size, for proficiency level ($F (2, 207) = 8.82, p = .001, \eta^2_p = .08$). 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 ($t = 2.01, p = .04$). 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, 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 ($F (2, 207) = 5.94, p = .005, \eta^2_p = .05$) and a main effect, and large effect size, for proficiency level ($F (2, 207) = 12.69, p = .001, \eta^2_p = .11$). NPANOVAs indicated 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 ($t = 2.91, p = .004$). 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-unit 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 a large effect size for task type ($F(2, 207) = 11.95$, $p = .001$, $\eta^2_p = .10$). 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 ($t = 3.47$, $p = .003$). In sum, 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 a main effect, and large effect size, for proficiency level ($F(2, 207) = 46.12$, $p = .001$, $\eta^2_p = .31$). 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 ($t = 4.60$, $p = .001$). 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 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 \((F(2, 207) = 10.01, p = .001, \eta_p^2 = .09)\). 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 \((t = 4.25, p = .001)\). The mean ratings for the 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 produced main effects, and moderate effect sizes, for task type \((F(2, 207) = 4.23, p = .002, \eta_p^2 = .04)\) and for ESL proficiency level \((F(2, 207) = 5.94, p = .005, \eta_p^2 = .05)\). 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 \((t = 2.64, p = .001)\). 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 \((t = 3.15, p = .001)\). 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 \((F(2, 207) = 14.27, p = .001, \eta_p^2 = .12)\). 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 \((t = 3.50, p = .003)\). The mean quality of data tended to increase by proficiency levels for the independent essays \((M = 0.8, M = 1.0, \text{ and } M = 1.1 \text{ for levels 3, 4, and 5, respectively})\), for the reading-writing tasks \((M = 0.7, M = 0.8, \text{ and } M = 1.8 \text{ for levels 3, 4, and 5, respectively})\), and for the listening-writing tasks \((M = 1.2, M = 1.8, \text{ and } M = 1.8 \text{ for levels 3, 4, and 5, respectively})\). 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 \(F (2, 207) = 10.47, p = .001, \eta^2_p = .09\). NPNANOVAs showed, for the quality of warrants in arguments, significant differences between the independent writing and reading-writing tasks \((t = 3.23, p = 0.04)\) and between the independent writing and listening-writing tasks \((t = 3.62, p = .001)\). There were few warrants utilized overall in the compositions, but these tended to be more prevalent in the independent essays \((M = 0.3, M = 0.2, \text{ and } M = 0 \text{ for levels 3, 4, and 5, respectively})\) than in the integrated tasks, where they scarcely appeared at all in either the reading-writing tasks \((M = 0, M = 0, \text{ and } M = 0.1 \text{ for levels 3, 4, and 5, respectively})\) or the listening-writing tasks \((M = 0, M = 0.1, \text{ and } M = 0 \text{ for levels 3, 4, and 5, respectively})\). 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 \(F (2, 207) = 24.70, p = .001, \eta^2_p = .19\). NPNANOVAs 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 \((t = 4.72, p = .001)\). 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 similar pattern emerged 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. 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, as described in Appendix A. 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.

The vast majority of T-units that examinees wrote in their compositions in the Field Test did not specify any source evidence, a tendency that was consistent across task types and proficiency levels. Consequently, the NPMANOVA did not produce any main effects. 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).

For the percentage of T-units in which a source was specified as the self, a nonparametric MANOVA showed a main effect, and large effect size, for task type ($F(2, 207) = 251.51, p = .001, \eta^2_p = .71$) and a main effect, but small effect size, for proficiency level ($F(2, 207) = 2.8, p = .03, \eta^2_p = .03$). 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 ($t = 18.78, p = .001$). 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 prompts for the tasks 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 ($F(2, 207) = 29.41, p = .001, \eta^2_p = .22$). 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 ($t = 2.75, p = .006$). 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 people 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 NPANOVAs showed, for the percentages of T-units in which source information was presented as a declaration, statistically significant differences 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 ($t = 16.43, p = .001$). 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.

**Orientations to Source Evidence: Message**

As explained in Appendix A, 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) a quotation, (c) a paraphrase, or (d) a summary. For the percentage of T-units in which source information was presented as a declaration, an NPANOVAs showed, for the percentages of T-units in which source information was presented as a declaration, statistically significant differences 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 ($t = 16.43, p = .001$). 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.
T-units in which source information was presented as a quotation were infrequent overall, and their appearance varied by task and by proficiency level. For this reason 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 ($F (2, 207) = 489.66, p = .001, \eta_p^2 = .83$). 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 ($t = 2.26, p = .006$). 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 ($F (2, 207) = 79.72, p = .001, \eta_p^2 = .44$) and a main effect, but small effect size, for proficiency level ($F (2, 207) = 4.54, p = .006, \eta_p^2 = .04$). 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 ($t = 12.42, p = .001$). For proficiency level, NPANOVAs showed 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, 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 by tallying 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 ($t = .92, p = \text{n.s.}$). 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., Politics produced more verbatim phrases for all groups than Cinema 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 2 and 3 highlight information related to verbatim uses of source texts, adding calculations of the percentages of verbatim words that appear in each composition, and comparing these by examinees’ proficiency levels. As shown in Table 2, 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 the less proficient (i.e., level 3) examinees’ compositions, 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 3 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 (a point we discuss below).
Table 2
Verbatim Phrases from Source Texts in the Two Reading-Writing Tasks

<table>
<thead>
<tr>
<th>Proficiency Level (by Independent Writing task)</th>
<th>Politics</th>
<th>Politics</th>
<th>Politics</th>
<th>Cinema</th>
<th>Cinema</th>
<th>Cinema</th>
<th>Both tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$ verbatim phrases</td>
<td>$M$ words verbatim from source</td>
<td>% of verbatim words in composition</td>
<td>$M$ verbatim phrases</td>
<td>$M$ words verbatim from source</td>
<td>% of verbatim words in composition</td>
<td>$M$ verbatim phrases</td>
</tr>
<tr>
<td>3 ($n = 12$)</td>
<td>11.2</td>
<td>53.7</td>
<td>33%</td>
<td>9.1</td>
<td>55.0</td>
<td>37%</td>
<td>10.1</td>
</tr>
<tr>
<td>4 ($n = 12$)</td>
<td>12.2</td>
<td>57.7</td>
<td>28%</td>
<td>5.0</td>
<td>23.3</td>
<td>16%</td>
<td>8.6</td>
</tr>
<tr>
<td>5 ($n = 12$)</td>
<td>11.7</td>
<td>48.7</td>
<td>22%</td>
<td>6.5</td>
<td>27.6</td>
<td>17%</td>
<td>9.1</td>
</tr>
</tbody>
</table>

Table 3
Verbatim Phrases from Source Texts in the Two Listening-Writing Tasks

<table>
<thead>
<tr>
<th>Proficiency Level (by Independent Writing task)</th>
<th>Plato</th>
<th>Plato</th>
<th>Plato</th>
<th>Behaviorism</th>
<th>Behaviorism</th>
<th>Behaviorism</th>
<th>Both tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$ verbatim phrases</td>
<td>$M$ words verbatim from source</td>
<td>% of verbatim words in composition</td>
<td>$M$ verbatim phrases</td>
<td>$M$ words verbatim from source</td>
<td>% of verbatim words in composition</td>
<td>$M$ verbatim phrases</td>
</tr>
<tr>
<td>3 ($n = 12$)</td>
<td>3.4</td>
<td>11.6</td>
<td>9%</td>
<td>3.2</td>
<td>10.3</td>
<td>11%</td>
<td>3.3</td>
</tr>
<tr>
<td>4 ($n = 12$)</td>
<td>7.1</td>
<td>24.8</td>
<td>13%</td>
<td>6.2</td>
<td>19.7</td>
<td>12%</td>
<td>6.6</td>
</tr>
<tr>
<td>5 ($n = 12$)</td>
<td>6.9</td>
<td>23.5</td>
<td>12%</td>
<td>5.5</td>
<td>19.1</td>
<td>13%</td>
<td>6.2</td>
</tr>
</tbody>
</table>
Impressions of Ineffective and Effective Writing for Each Task Type

To exemplify the discourse features we coded, we describe below some impressions of qualities of writing that seemed to be particularly ineffective and effective for each of the three task types, as represented by compositions that received exceptionally low or high ratings on the tasks. We attend particularly to the examinees’ functional uses of evidence, argumentation, and source materials in these compositions. We caution, however, that the compositions were characterized by variability rather than uniformity, so our observations here are highly interpretive and not meant to be representative descriptions of writing typical of any particular score levels on the test.

Independent Essays. Ineffective independent essays tended to present a clearly stated proposition and to make several claims, but often did not provide data to support the claims. Transitional phrases tended to be lacking to mark the argument structure or its conclusion. Such writers made numerous grammatical, spelling, and punctuation errors, used a simple range of vocabulary, and displayed little variety in sentence structure. There tended to be conflicts in orientation. For example, some writers began some sentences with “I think”, presenting the self as the primary source of evidence, but then made declarations in the second person (“you” or “your”), as if giving advice to the reader, or also in the third person (“they” or “people”), as if referring to other people.

The most proficient writers in our sample tended to produce independent essays with a single extended argument, developed creatively through numerous claims and supporting data. They marked their argument structure by a variety of transition phrases in paragraphs of varying sizes, but not in a manner that seemed rhetorically formulaic (as compositions at the mid-range often tended to do). Their discussion adopted either an academic tone, using third person throughout, or a personal tone, using first person consistently. Apart from minor spelling errors, their range of vocabulary was varied and precise; their language forms were essentially native-like with many complex embedded syntactic structures and nominal or adverbial phrases. For instance, rather than common
phrases like “I think” to express the self as the primary source of evidence, one examinee stated “I am of the view that…”

**Integrated Reading-Writing Tasks.** The least proficient students in the sample tended to write compositions for the integrated reading-writing task that had introductory paragraphs that alluded to supporting points. But either (a) their claims supporting the opening proposition were not developed in the composition or (b) the examinees tended to dwell solely on a single personal experience of limited relevance to the main proposition. Varied points appeared, giving an impression of incoherence. Sentences were structured erratically, for instance, with long, run-on sentences, sentence fragments, or other inappropriate uses of punctuation. Minor spelling and word-form errors abounded. The vocabulary was limited and occasionally incomprehensible. Some examinees used the self as a source of evidence (e.g., “I can say that…”), which seemed inappropriate for this task. Verbatim phrases from the source text appeared frequently in these ineffective compositions, but there were seldom any direct acknowledgements of the source reading material.

The most proficient examinees in our sample were 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. (The mid-range compositions on this task tended to paraphrase ideas from the source readings, point by point, using many verbatim phrases from the source reading, as we described above.) Where phrases in the particularly effective compositions may have derived from the source text, they were often common expressions in ordinary usage (e.g., “at the same time”) that were difficult to consider as textual borrowing. Some highly effective responses to the reading-writing task were relatively brief, and some even lacked a developed paragraph format, giving them the quality of a text that may appear in, for example, an exam response in a university course. A main proposition tended to be stated at the beginning and reiterated at the end of these compositions, supported by concise claims and relevant data. The choice of words seemed original and precise, adopting an academic register. Correspondingly, most T-units were declarations in the third person, summarizing the source text if not always citing it as the source of evidence. These texts were written in a
variety of syntactic structures, utilizing a range of verb tenses, with only a few minor errors.

**Listening-reading tasks.** Many of the least proficient examinees in our sample tended to phrase their listening-reading tasks in a structured, formulaic manner (e.g., “First,” “Secondly,” “Finally”). However, their introductions often failed to address the question posed in the task prompt, and their conclusions were often difficult to perceive if evident at all. The stated propositions and claims were weak, often containing facts whose relevance to the task was not clear. Few of the least proficient examinees used verbatim phrases from the source text, perhaps because they did not comprehend or could not recall hearing them. Indeed, some of these ineffective compositions referred to information that had not, in fact, appeared in the source text. Their uses of English were characterized by simple sentences, clauses, and vocabulary.

Particularly effective compositions for the listening-writing tasks in our sample tended to state a main proposition in both their introductory and concluding statements, presenting in the main text an extended discussion of the proposition, linked by transition phrases between paragraphs. Rather than recycling chunks of data from the lecture (as the mid-range or ineffective compositions tended to do), the effective responses on this task made specific, well-connected claims related to the main proposition. Notably, as in the reading-writing task, the most proficient examinees summarized, rather than paraphrased or used verbatim phrases from, the source text. Their language contained minor errors, many of which may have been typographical in origin, but also contained a variety of linguistic structures, succinct clauses, and appropriate lexical choices. They tended to acknowledge the source text explicitly (e.g., “In this book, The Republic, Plato explains…”).

**Discussion and Implications**

There were statistically 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 Next Generation
TOEFL. The results of NP MANOVAs 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., for 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 (and thus
verifying that they are noteworthy indicators of development in writing ability in ESL).
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 in 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, 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 oneself, 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 Next Generation 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 nor 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 (reading-writing vs. listening-writing) tasks. This was true for most of the features of argument structure (which did not feature much in the integrated tasks) and for examinees’ distinctions between evidence arising from the self or from a source text. The two types of integrated tasks appear, in these respects, to prompt examinees to produce written discourse that shares common features.

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 Test of Written English) as well as for those being developed for the integrated tasks for Next Generation 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, piece-meal 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 Next Generation TOEFL 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 Next Generation TOEFL. 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 academic-type source texts. The integrated tasks require complex cognitive, literate, and language abilities for comprehension as well as to produce 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 present 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. As such, 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 qualities of written discourse structure.

Likewise, orientation and instructional materials should be useful, as in the LanguEdge Courseware (Educational Testing Service, 2002, and see new materials being produced at http://toefl.org/imaterials.html), to facilitate examinees preparing to write integrated reading-writing and listening-writing tasks on Next Generation TOEFL. The present study points toward important differences between examinees being able (a) to paraphrase bits of information from source texts or (b) to 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 article. Orientation and instructional materials should be able to help examinees prepare to write integrated tasks appropriately on the test, but more importantly, 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 Next Generation 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


**Appendix A. 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>
<thead>
<tr>
<th>ID</th>
<th>Voice</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unspecified</td>
<td>Nowadays children do not like mathematics. Obviously, economy is getting worse.</td>
</tr>
<tr>
<td>2</td>
<td>Self</td>
<td>I promise I won’t keep you a moment longer. I think he was a bit shorter than you are.</td>
</tr>
<tr>
<td>3</td>
<td>Specified other(s)</td>
<td>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.</td>
</tr>
<tr>
<td>4</td>
<td>Unspecified other(s)</td>
<td>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.</td>
</tr>
<tr>
<td>5</td>
<td>Community</td>
<td>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.</td>
</tr>
</tbody>
</table>

**Decision tree for categorizing “voice”**

```
Voice
  / \
unspecified specified
  / \
self others
  / \
identified not identified
  (specified others) (unspecified other(s) or community)
```

**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 (1996) definitions of paraphrase and summary. Thompson (1996) 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, i.e., 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>
<thead>
<tr>
<th></th>
<th>Declaration</th>
<th>No reference to a source text. Statement of personal opinion or fact.</th>
<th>I think most people are nice. Most people are nice.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Quotation</td>
<td>Verbatim quotation from source text set off by quotation marks.</td>
<td>‘Why are you not Orthodox?’ people say. Finally he lifted his chin and spoke. ‘I could swim when I was five.’</td>
</tr>
<tr>
<td>2</td>
<td>Paraphrase</td>
<td>The writer paraphrases one idea from the source.</td>
<td>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.</td>
</tr>
<tr>
<td>3</td>
<td>Summary</td>
<td>The writer summarizes or makes a generalization about two or more ideas from the source text.</td>
<td>These three groups have different roles in the society.</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Independent Writing Task Scoring Guidelines (from *LanguEdge Courseware*, 2002, p. 35, Educational Testing Service, cf. TOEFL Essay or Test of Written English)

**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

Reading/Writing Task Scoring Guidelines (from LanguEdge Courseware, 2002, p. 47, Educational Testing Service)

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 more 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)

Listening/Writing Task Scoring Guidelines (from LanguEdge Courseware, 2002, p. 61, Educational Testing Service)

5 A response at this level
• amplly 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).