Article

LANGUAGE TESTING

Language Testing
2016, Vol. 33(3) 319–340

DOI: 10.1177/0265532215587391
ltj.sagepub.com
SAGE

Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing

Kristopher Kyle
Georgia State University, USA

Scott A. Crossley
Georgia State University, USA

Danielle S. McNamara
Arizona State University, USA

Abstract

This study explores the construct validity of speaking tasks included in the TOEFL iBT (e.g., integrated and independent speaking tasks). Specifically, advanced natural language processing (NLP) tools, MANOVA difference statistics, and discriminant function analyses (DFA) are used to assess the degree to which and in what ways responses to these tasks differ with regard to linguistic characteristics. The findings lend support to using a variety of speaking tasks to assess speaking proficiency. Namely, with regard to linguistic differences, the findings suggest that responses to performance tasks can be accurately grouped based on whether a task is independent or integrated. The findings also suggest that although the independent tasks included in the TOEFL iBT may represent a single construct, responses to integrated tasks vary across task sub-type.

Keywords

Integrated tasks, language use domain, natural language processing, speaking assessment, TOEFL iBT

Corresponding author:
Kristopher Kyle, Department of Applied Linguistics/ESL, 34 Peachtree St. Suite 1200, One Park Tower Building, Georgia State University, Atlanta, GA 30303, USA.
Email: kkyle3@student.gsu.edu

Introduction

The last two decades have seen an increased interest in direct assessments of language proficiency (e.g., Camp, 1993; Chapelle, Enright, & Jamieson, 2008; Hamp-Lyons, 1991; Weigle, 2010). Such interest has led to the overhaul of influential tests such as the Test of English as a Foreign Language (TOEFL; see, e.g., Chapelle et al., 2008 for a discussion of these changes). Traditionally, direct assessments of language proficiency have included *independent* tasks, or impromptu tasks that do not require test-takers to read or listen to an input sample in order to complete the task. It has been argued that such tasks, although prevalent, make unwarranted assumptions about a test-taker's background knowledge (Read, 1990). In addition, it has been argued that independent tasks under-represent speaking proficiency and have been criticized for a lack of authenticity (Barkaoui, Brooks, Swain, & Lapkin, 2013; Brown, Ishiwata, & McNamara, 2005). As such, *integrated* tasks, which require test-takers to read and/or listen to a language sample and integrate the content of those samples into a response, have been added to tests such as the Test of English as a Foreign Language internet-Based Test (TOEFL iBT). Researchers have argued that integrated tasks more accurately and effectively measure test-taker's language abilities and improve washback effects (Cumming et al., 2005). Furthermore, the inclusion of multiple task types may strengthen the *explanation inference* of a test validity argument (i.e., the inferences that can be made about test-taker ability based on scores; Chapelle, Enright, & Jamieson, 2008) by modeling more authentically the various types of tasks that are often encountered in academic settings.

In light of perceived differences between independent and integrated tasks, it becomes important to establish that the two tasks elicit different types of language performance in assessment situations (i.e., in the TOEFL; Cumming et al., 2005). In the context of the writing portion of the TOEFL iBT, linguistic differences between performances on independent and integrated writing tasks have been documented by using computer systems (i.e., natural language processing: NLP) to analyze textual features automatically (e.g., Guo, Crossley, & McNamara, 2013; Cumming et al., 2005). These differences are less well documented for the tasks included in the speaking portion of the TOEFL iBT. Brown et al. (2005) explored linguistic differences between five prototype speaking tasks that were considered during the construction of the TOEFL iBT. The linguistic indices that were used in the study related to grammatical accuracy and complexity, word frequency, phonological accuracy, and text content. Brown et al. (2005) found few (and small) differences between integrated and independent speaking tasks. More recently, Biber and Gray (2013) explored lexico-grammatical features of TOEFL iBT independent and integrated written spoken and written task types that appear in current versions of the test. One limitation of the Biber and Gray study is that it combined the two independent task types included in the TOEFL iBT into a single construct of independent speaking and the four integrated task types included in the TOEFL iBT into a single construct of integrated speaking. This is potentially problematic because the tasks differ not only with regard to the use of source material (i.e., no source material, listening passage, reading and listening passage), but also with regard to language use domain (e.g., campus situation versus academic course).

This study builds on Brown et al. (2005) and Biber and Gray (2013) by examining responses to operational TOEFL iBT speaking tasks using linguistic indices related to lexical characteristics (i.e., lexical sophistication and semantic and lexical categories) and cohesion without treating the speaking tasks as interchangeable. Lexical and cohesion features have been shown to be important predictors of speaking and writing quality in a number of assessment contexts (Crossley & McNamara, 2013; Guo et al., 2013; Crossley, Clevinger, and Kim 2014). Furthermore, lexical and cohesion features align with the expectations found in rating scales used for the TOEFL speaking tasks. The objective of this study is to distinguish not only independent from integrated tasks but also to distinguish between independent tasks types and integrated task types, which differ with regard to language use domain. In doing so, we attempt to clarify the construct validity arguments for the inclusion of various tasks found in the speaking section of the TOEFL iBT. Such arguments have not been addressed before with regard to the lexical and cohesive characteristics of responses to individual task types and, as such, carry important implications for language assessment, particularly regarding research methodology, item writing, and item assessment.

Independent and integrated writing

The majority of research on differences between independent and integrated tasks has occurred in the context of writing assessment. For instance, Cumming et al. (2005) investigated differences in responses to Test of Written English (TWE)/TOEFL independent essays and two prototype integrated writing tasks (read/write and listen/write). Cumming and colleagues found that essays written in response to integrated tasks tended to be shorter, included more word types, included longer T-units with more clauses, be less argumentative, and referenced and relied on outside source more often than responses to independent essay tasks. Because the findings indicated that each task led to the production of diverse linguistic features, and therefore likely assessed distinct writing constructs, they argued for the inclusion of both independent and integrated tasks in the TOEFL iBT.

Guo, Crossley & McNamara (2013) built on the findings of Cumming et al. (2005) by investigating differences between operational TOEFL iBT independent and integrated written tasks. Guo et al. used advanced NLP tools to evaluate textual characteristics and to create statistical models that indicated the relative importance of these characteristics for predicting independent and integrated essay scores. They used Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Graesser, McCarthy, & Cai, 2014) variables related to text length, lexical sophistication, syntactic complexity, and cohesion to build models of writing quality for both independent and integrated essays. They found that 19 unique (non-collinear) text variables significantly correlated with integrated essay scores, and that 21 unique text variables correlated with independent essay scores. For the integrated texts, seven of these variables were chosen by a stepwise multiple regression that accounted for 53.3% of the variance in scores. For independent essays, five of the variables with significant correlations to independent essay scores were chosen by a stepwise multiple regression that accounted for 57.4% (test set) of the variance in the essay scores. Only two variables overlapped between the two models

(number of words per text and past participle verbs). Guo et al., interpreted these findings as providing support for the inclusion of both independent and integrated writing tasks because the regression models for the two tasks differed in the types of linguistic features that were predictive of essay quality across the two tasks.

Taken together, the studies by Cumming et al. (2005) and Guo et al., 2013 (2013) demonstrate the usefulness of employing textual analyses when evaluating language performance tasks. Additionally, they demonstrate the efficacy of employing advanced NLP tools and confirmatory statistics such as multiple regression to determine the relative importance of textual differences.

Independent and integrated speaking

Brown et al. (2005) published an in-depth study on the development of the TOEFL iBT speaking tasks. The study included an examination of rater-cognition and a linguistic analysis of responses to five in-development tasks. The prototype speaking tasks used in the Brown et al. study included two independent prompts that asked test-takers to provide their opinion on a particular topic. The speaking tasks also included an integrated task in which the test-takers listened to a monologic lecture and then were asked to respond to a related question, a similar integrated task in which the listening sample was dialogic in nature, and an integrated task in which test-takers read a passage and then responded to a related prompt. The rater cognition study, which asked experienced ESL instructors to provide feedback on the quality of task response samples and elaborate on their reasons for judging a particular sample to be of high or low quality, formed the basis for the current independent and integrated rubrics used in the scoring of TOEFL iBT speaking tasks.

The linguistic analysis found in the Brown et al. (2005) study sought, among other things, to determine whether there were differences in performance across proficiency levels and task types. To do so, a number of linguistic analyses were conducted to explore the production of features identified in the rater-cognition study (linguistic resources, content, phonology, and fluency). For these analyses, the Computerized Language Analysis (CLAN) program (MacWhinney, 2000) was used in addition to phonetic analyses conducted by a phonetician. The results of ANOVA statistical tests indicated that although there were a number of differences between independent and integrated tasks, only four measures achieved an effect size that was larger than "marginal". These effects demonstrated that responses to independent tasks included more words from the 1K list (the 1000 most frequent words found in a corpus of English), a higher number of modal verbs, more word-types, and a longer mean length of run (the average number of syllables per utterance) than responses to integrated tasks. Based on these findings, the authors concluded that "there appears to be little evidence to support the view that integrated task performance will be more linguistically complex or sophisticated than on independent tasks" (pp. 105–106). One must be cautious, however, when attempting to generalize the results of Brown et al. (2005) to operational TOEFL speaking tasks, because only one of the prototype integrated tasks (monologic listen/speak task) is similar to the operational tasks currently found in the TOEFL iBT.

Lee (2006) also investigated the prototype speaking tasks found in Brown et al. (2005), but focused on the reliability of a number of test configurations. Lee investigated the effects of the inclusion of 1–12 speaking tasks rated by 1–2 raters. He found that five tasks rated by a single rater each provided the most advantageous configuration that balanced reliability and cost. Based on the findings of Lee (2006) and Brown et al. (2005), in addition to other testing factors, the final version of the speaking section of the TOEFL iBT includes two independent tasks, two read/listen/speak integrated tasks, and two listen/speak integrated tasks (see Wang, Eignor, & Enright, 2008 for more details).

Building on Cumming et al. (2005) and Brown et al. (2005), Biber and Gray (2013) explored the lexico-grammatical characteristics of operational TOEFL iBT integrated and independent tasks for both spoken and written modes and across score levels. To do so, they employed vocabulary distributions such as the percentage of words in the target text from the 1K and 2K (the second most frequent words in a corpus of English texts) word lists, the Academic Word List (AWL; Coxhead, 2000), collocational patterns with common verbs, lexical bundles (Biber, Conrad, & Cortes, 2004), and grammatical and lexico-grammatical patterns measured by the Biber Tagger (Biber, 1988, 1995).

In a full factorial model that explored lexico-grammatical features across task type (independent vs. integrated), mode (spoken vs. written) and score, Biber and Gray identified 36 features that met their significance threshold (see Biber & Gray, 2013, pp. 41–44 for an overview of these results). Particularly pertinent to the current study was a multidimensional analysis (MDA; Biber, 1988), which examined the lexico-grammatical features identified in the full factorial model. Four dimensions were identified and interpreted as *literate versus oral*, *information source: text versus personal experience*, *abstract opinion versus concrete description/summary*, and *personal narration*. After mapping the mean score for each text type and task score (e.g., samples that were written, integrated, and earned a task score of 4) to each dimension, Biber and Gray (2013) found that written integrated responses were scored as more literate, whereas spoken independent responses were scored as more oral. Written integrated responses scored highest on the information source dimension followed by spoken integrated responses, written independent responses, and spoken independent responses. On the abstract opinion versus concrete description/summary dimension, independent written responses tended to be the most abstract. Written integrated responses were also more abstract than any of the spoken responses (independent or integrated), which tended to be highly concrete. The last dimension indicated that spoken independent responses had the most personal narration features, followed by written independent responses, spoken integrated responses, and then written integrated responses. Biber and Gray (2013) concluded that responses to TOEFL iBT independent and integrated tasks in both spoken and written modes differ in a number of ways that align with previous investigations into academic written and spoken discourse.

Although Biber and Gray's (2013) results provide information about differences between independent and integrated tasks, their division of tasks solely as either independent or integrated could lead to overgeneralizations. Although the two independent speaking tasks differ only slightly with regard to prompt type, the four integrated speaking tasks differ in potentially meaningful ways with regard to language use domain (campus situation versus academic course) and type of input that test-takers

receive (a listening passage alone versus a listening passage and a reading passage). Because Biber and Gray (2013) made no indication as to whether language use domain or type of input was controlled for, it remains unclear whether the characteristics of the dimension loadings are representative of each task. Furthermore, no indication was made that the authors controlled for test form (two forms of the TOEFL iBT comprised their corpus), so interpretations of whether the dimension loadings are characteristic of responses of all independent or integrated responses in each mode (i.e., spoken and written) should be made with caution. Finally, before firm conclusions can be made regarding the textual, linguistic features of TOEFL iBT speaking tasks, confirmatory statistical models (e.g., discriminant function analysis) should be conducted to determine whether features identified in the exploratory MD can be used to accurately predict the group membership (task type and/or mode) of a particular text (e.g., Crossley, Clevinger, & Kim, 2014).

Crossley et al. (2014) explored the lexical and cohesive features of a TOEFL integrated listen/speak task and responses to that task. The study had two focuses. First, Crossley et al. investigated which lexical and cohesive features predicted whether words in the listening passage would be repeated in test-taker responses. Second, Crossley et al. (2014) investigated the whether these lexical and cohesive features as found in test-taker responses predicted scores given to those responses by expert raters. They found that lexical and cohesive features such as the repetition of particular words in the source text were predictive of whether those words would be repeated or not and also predictive of speaking proficiency scores. They also reported that a number of lexical and cohesive features were correlated with the scores assigned to responses. Hypernymy scores for words not included in the source text, for example, were negatively correlated with holistic scores, suggesting that responses containing words that have more hypernymic relationships (i.e., are more specific) earn higher scores. Although all the indices included in Crossley et al.'s (2014) predictor models were prompt-specific, the study highlights the importance of indices of lexical sophistication and cohesion in TOEFL speaking responses.

Barkaoui et al. (2013) represent a departure from investigations of task responses by examining the strategies employed by test-takers to complete the tasks. They found that although differences existed in strategy use between all tasks, the strategies used by test-takers when responding to each of the integrated tasks were more similar to each other than those used when responding to the independent tasks. For example, test-takers used more strategies when responding to integrated tasks than when responding to independent tasks. These results contribute to the validity argument for the inclusion of both independent and integrated tasks.

The extant research provides some evidence that the TOEFL iBT independent and integrated spoken tasks require different strategies (Barkauoi et al., 2013) and elicit responses with different textual features (Biber & Gray, 2013). What is not clear, however, is whether the presence of source material in the task is the only factor that affects the linguistic characteristics of responses. The target language use domain of a particular task (i.e., a campus situation versus an academic course), for example, may affect the type of language produced in response to a particular task. Additionally, previous research has not explored whether the linguistic differences between task responses are

strong enough to distinguish accurately between task types. This study attempts to address these gaps by responding to the following research questions:

1. Are there differences in test-taker responses to operational TOEFL iBT speaking tasks with regard to lexical characteristics and cohesive features?
2. If such differences exist, can they be used to classify responses accurately according to task type?

Method

Corpus

The current study examines a subset of the public-use TOEFL speaking samples provided by Educational Testing Service (ETS). These samples include responses produced by test-takers during actual administrations of two forms of the TOEFL iBT in 2007. The corpus is composed of a stratified random sample that included a balanced number of responses from two independent tasks and two integrated tasks that earned scores of "3" or "4". In total, 240 samples were used and were balanced with regard to task, form, and scores.

In the independent tasks, test-takers are provided with a prompt and then given 15 seconds to prepare their response. After the preparation time, test-takers are given 45 seconds to respond to the prompt. The first task asks test-takers to "express and defend a personal choice from a given category" (ETS, 2012, p. 15). These categories can include "important people, places events or activities you enjoy" (p. 15). The second task asks test-takers to "make and defend a personal choice between two contrasting behaviors or courses of action" (p. 15). Both independent tasks are rated using the same 4-point rubric, which includes three key areas: delivery, language, and topic development. The rubrics for the independent tasks can be found at www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf.

The TOEFL iBT includes four integrated tasks. The first two include both a reading and a listening passage, while the second two include only a listening passage. In the read/listen/speak task, test-takers have access to the reading passage while composing their response, affording textual borrowing that could influence the structure of their response. Since part of our interest lies in the linguistic structure produced by the test-taker (i.e., the cohesion of the response), we focus only on the two independent and the two listen/speak tasks and ignore the read/listen/speak tasks. Such an approach allows us to examine better the natural language produced by test-takers. During the listen/speak tasks, test-takers listen to a listening passage and then are provided with a prompt and given 20 seconds (tasks 5 and 6) to prepare a response. After the preparation time, test-takers are given 60 seconds to deliver their response. The two listen/speak tasks comprise a campus situation prompt and an academic course prompt. The campus situation task asks test-takers to listen to a conversation about a problem one of the speakers is having, and then "to demonstrate an understanding of the problem and to express an opinion about solving the problem" (ETS, 2012, p. 15). The academic course task asks test-takers to "summarize the lecture and demonstrate an understanding of the

relationship between the examples and the overall topic" (p. 15). A single 4-point rubric is used to evaluate the integrated tasks, and includes descriptors concerning delivery, language use, and topic development. Although these general descriptor categories are the same in the rubric for the independent tasks, the specific descriptors are slightly different.

Because the focus of the present study is on the differences and similarities between TOEFL iBT speaking task types, only responses that were given a "3" or "4" by ETS raters were considered. This score level can be considered to represent *successful* completion of a particular task, and therefore is an indication that test-takers understood and addressed the task appropriately. All the spoken responses were orthographically transcribed by at least two trained transcribers. Differences in transcriptions were adjudicated by a third transcriber.

Text analysis

The current study uses three NLP tools: the Tool for the Automatic Analysis of Lexical (TAALES; Kyle & Crossley, 2015), Coh-Metrix (Graesser, et al. 2004; McNamara et al., 2014) and lists from Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) to analyze the linguistic features of each response. All three of these tools have been used successfully to analyze linguistic output in assessment settings. TAALES includes a broad range of lexical sophistication indices and has recently been used to evaluate responses TOEFL independent speaking tasks (e.g., Kyle & Crossley, 2015). Coh-metrix includes a broad range of cohesion indices and has been used to investigate a number of integrated and independent TOEFL writing and speaking tasks (e.g., Guo et al., 2013; Crossley et al., 2014). LIWC indices comprise a number of grammatical and semantic features that are lexically based (e.g., involve no POS tagging), and therefore are reliable for analyzing spoken data. LIWC indices have been successfully used to analyze independent spoken responses (e.g., Crossley & McNamara, 2013).

TAALES. TAALES is an automatic text analysis program that was designed to investigate lexical sophistication. This study included 114 TAALES indices related to lexical frequency (i.e., the number of times an item occurs in a reference corpus), lexical range (i.e., the number of documents in a reference corpus in which an item occurs in), psycholinguistic word information (e.g., concreteness and meaningfulness), and academic language (i.e., items that occur more frequently in an academic corpus than in a general use corpus) for both single words and multi-word units (n-grams). The frequency and range indices draw on the British National Corpus (BNC; 2007), Thorndike–Lorge Corpus (Thorndike & Lorge, 1944), Brown corpus (Kucera & Francis, 1967), Brown verbal frequencies (Brown, 1984), which were compiled based on the London–Lund corpus of English Conversation (Svartvik & Quirk, 1980), and the SUBTLEXus corpus of subtitles (Brysbaert & New, 2009). Psycholinguistic word information indices draw on the Medical Research Council (MRC) psycholinguistic database (Coltheart, 1981), which includes word scores for familiarity, concreteness, imageability, and meaningfulness. Also included in TAALES are psycholinguistic word information indices that incorporate recently collected concreteness norms for single words and two-word units (Brysbaert,

Table 1. Overview of features considered.

Bank of indices	TAALES	Coh-Metrix	LIWC	Total
Word frequency				
Written	18			18
Spoken	18			18
Combined		3		3
N-gram frequency				
Written	8			8
Spoken	8			8
Word range				
Written	9			9
Spoken	9			9
N-gram proportion scores				
Written	2			2
Spoken	2			2
Academic language	15			15
Psycholinguistic information				
Words	21			21
Bigrams	1			1
Combined	3			3
Fluency		2		2
Cohesion		24		24
Readability		2		2
Word information		3		3
Part of speech		15	18	33
Semantic categories			21	21
Total	114	49	39	202

Warriner, & Kuperman, 2014) and age of acquisition norms for single words (Kuperman, Stadthagen-Gonzales, & Brysbaert, 2012). Academic language frequencies are based on the Academic Word List (AWL; Coxhead, 2000) and the Academic Formulas List (Simson-Vlach and Ellis, 2010). For an overview of the TAALES index categories used in this study, see Table 1.

Coh-Metrix. Coh-Metrix is an automatic text analysis tool designed to investigate textual cohesion and coherence. The online version, which is available at cohmetrix.com includes indices that measure various textual features, including descriptive indices (number of words, number of paragraphs, etc.), text easability (e.g., McNamara et. al., 2011), referential cohesion (e.g., noun overlap), latent semantic analysis (a vector model method of comparing the semantic similarity of two texts or text segments, such as sentences or paragraphs), lexical diversity, incidence of connectives (e.g., causal connectives), situation model (e.g., causal verb incidence), syntactic complexity (e.g., number of modifiers before a noun phrase), syntactic pattern density (e.g., noun phrase density), word information (e.g., frequency, hypernymy), and readability (e.g., Flesch reading ease). Many of the

indices in Coh-Metrix employ a part of speech (POS) tagger and/or a syntactic parser. Because POS taggers and syntactic parsers are less reliable with spoken texts (see Biber & Gray, 2013 for a treatment of this), we omitted any indices that used such tools. Additionally, a number of indices included in Coh-Metrix are also included in TAALES. Where indices were similar, the TAALES version was used. Following these criteria, 47 Coh-Metrix indices of cohesion, word information, and lexical diversity were used. Additionally, two fluency indices (words per second and syllables per second) were adapted from existing Coh-Metrix indices, for a total of 49 Coh-Metrix indices. For an overview of the Coh-Metrix indices included in this study, see Table 1.

LIWC. LIWC includes 53 indices that measure basic text information (e.g., word counts), proxies for word frequency (e.g., number of words over six letters in length), POS counts (e.g., articles, prepositions, present tense verbs), and semantic categories (e.g., positive emotions, cognitive mechanisms, swear words). LIWC is distinct from other programs that include POS information, such as Coh-Metrix, in that all LIWC indices are lexically based, and therefore do not use a POS tagger or syntactic parser. In this study, we included 39 LIWC POS and semantic category lists. A number of LIWC indices were excluded based on construct irrelevance (e.g., swear words). For an overview of the LIWC index categories included in this study, see Table 1.

Statistical analyses

To investigate the research questions posed, we conducted two discriminate function analyses (DFA), following the same methods of variable selection. The first DFA investigated the differences between task types (i.e., independent versus integrated tasks) following previous research. The second DFA investigated the differences between the four tasks to determine the effects of language use domain and task type on linguistic production. Prior to inclusion in the analysis, all variables were first checked for normality, and any variables that violated normality were removed. To control for prompt differences (e.g., Hinkel, 2002), we examined the relative strength of the differences between prompts for each task type and the differences between task types (or grouped task types) via the reported partial eta squared effect sizes. If the differences for a variable were stronger between task types as compared to prompts, the variable was retained in the analysis.

We used multivariate analyses of variance (MANOVAs) to examine differences between remaining variables and the task types. Within these MANOVAs, we also checked for homogeneity of variance. If any variables failed to meet this assumption, they were removed from further analysis. The variable with the strongest effect size from each of the banks of indices was then preliminarily chosen for inclusion in the DFA. These variables were checked for multi-collinearity to ensure that each index was measuring a distinct construct and to avoid overfitting the DFA model. If any two variables were correlated at $r \geqslant .900$ (Tabachnick & Fidell, 2001), the variable with the largest difference between task types (measured by effect size) was kept. The remaining variables were then entered into a stepwise DFA. A follow-up DFA with leave-one-out

cross-validation (LOOCV) was used to ensure generalizability of the results across the entire dataset. LOOCV is a type of validation wherein a predictor model is created using all data points but one, and then tests the model on the remaining data point. This process is repeated until a group membership prediction has been made for all data points.

Results

Study 1: Combined independent tasks vs. combined integrated tasks

The first analysis investigated the degree to which responses to the combined independent tasks and the combined integrated tasks could be accurately classified. Of the 123 variables that were normally distributed, 15 showed larger differences between task type (combined independent tasks vs. combined integrated tasks) than between any of the differences between prompts for each task. After checking for homogeneity of variance, choosing the strongest predictor from each construct and controlling for multicollinearity, nine variables were entered into a stepwise DFA and a stepwise LOOCV DFA.

The model created by the stepwise DFA achieved a classification accuracy of 87.9% using five variables (*Givenness*, *Type-token ratio*, *Spoken range content words*, *Spoken bigram frequency*, and *Meaningfulness*; see Table 2 for a detailed description of these variables), which is significantly higher ($df = 1$, $n = 240$, $\chi^2 = 139.186$, $p < .001$) than the baseline accuracy of 50%. The reported Kappa = .758, indicates substantial agreement between actual and predicted task type (Landis & Koch, 1977). The stepwise LOOCV DFA achieved a classification accuracy of 87.5%, suggesting that the predictor model is stable across the dataset. Table 3 includes the descriptive and MANOVA statistics for the variables included in the DFA model.

Table 4 includes the confusion matrix for the stepwise DFA. The rows in the confusion matrix indicate how the actual responses to a particular task were classified. The columns indicate how often responses to tasks were classified as a particular task. Task-type group membership predictions were accurate for 92.5% of the combined independent task responses and 83.3% of the combined integrated task responses. The confusion matrix shows that 111 of the responses to the combined independent tasks were correctly classified, while nine of these responses were incorrectly classified as responses to combined integrated tasks. The confusion matrix also shows that 100 of the responses to the combined integrated tasks were correctly classified, while 20 of these responses were incorrectly classified as responses to combined independent tasks.

Table 5 comprises a fine-grained confusion matrix, which indicates how responses to each subtask (Independent – Personal Preference, Independent Choice, Integrated – Campus Situation, and Integrated– Academic Course) were classified (either as combined independent or combined integrated). These results indicate that responses to both independent subtasks were correctly categorized with accuracy above 90%. Additionally, responses to the Integrated – Academic Course subtask were correctly classified as combined integrated with an accuracy of 95%. Responses to the Integrated – Campus Situation subtask, however, were only correctly classified with an accuracy of 71.7%.

Table 2. Description of indices included in the combined independent vs. combined integrated model.

Variable	Variable code	Source	Variable description
Spoken bigram frequency	BNCSpokenBigramNormedbiFreqLog	TAALES	The mean bigram frequency based on the spoken portion of the British National Corpus, log transformed
Givenness	LSAGN	Coh-Metrix	The ratio of given to new information as measured by latent semantic analysis
Meaningfulness	MRCMeaningfulnessAW	TAALES	The average meaningfulness value of all words, based on the Medical Research Council database
Spoken range content words	SUBTLEXusRangeCWLog	TAALES	The average range of content words based on the SUBTLEXus corpus of subtitles, log transformed
Type-token ratio	TYPETOKN	Coh-Metrix	The number of unique types divided by tokens

Table 3. Means (standard deviations), *F* values, and effect sizes for combined independent and combined integrated responses.

Variables	Combined independent	Combined integrated	*F*(1, 238)	η^2_p
Givenness	0.234 (0.068)	0.285 (0.055)	40.392	0.145
Type-token ratio	56.772 (7.130)	50.560 (5.672)	55.775	0.190
Spoken range content words	3.554 (0.108)	3.350 (0.184)	109.505	0.315
Spoken bigram frequency	2.735 (0.126)	2.673 (0.112)	15.849	0.062
Meaningfulness	409.247 (20.681)	414.719 (19.003)	41.445	0.148

Note: For all indices, $p < .001$.

Table 4. Confusion matrix for Study 1 DFA: Combined independent vs. combined integrated.

Actual task type	Predicted as combined independent	Predicted as combined integrated	Total
Combined independent	**111**	9	120
Combined integrated	20	**100**	120
Accuracy	92.5%	83.3%	**87.9%**

Table 5. In-depth confusion matrix for Analysis 1 DFA: Independent and integrated.

Actual task type	Predicted as combined independent	Predicted as combined integrated	Total	Accuracy
Independent personal preference	**56**	4	60	93.3%
Independent choice	**55**	5	60	91.7%
Total independent	**111**	9	120	92.5%
Integrated campus situation	17	**43**	60	71.7%
Integrated academic course	3	**57**	60	95.0%
Total integrated	20	**100**	120	83.3%
Total accuracy:				**87.9%**

Study 2: Independent – Personal Preference tasks vs. Independent – Choice tasks vs. Integrated – Campus Situation tasks vs. Integrated – Academic Course tasks

A second analysis investigated whether responses to the four subtask types (Independent – Personal Preference, Independent – Choice, Integrated – Campus Situation, and Integrated – Academic Course) could be accurately classified. Of the 123 variables that were normally distributed, 30 showed stronger differences between task types than between any of the differences between prompts for each subtask type. After checking for homogeneity of variance, choosing the strongest predictor from each construct, and

Table 6. Description of indices included in the four task-type predictor model.

Variable	Variable code	Source	Variable description
Insight words	Insight	LIWC	The mean number of *insight* words such as think, know, and consider
Personal pronouns	Ppron	LIWC	the mean number of personal pronouns
Motion prepositions	SPATmpi	Coh-Metrix	The number of motion prepositions per 1,000 words
Spoken range	SUBTLEXusRangeCWLog	TAALES	The average range of content words based on the SUBTLEXus corpus of subtitles, log transformed
Type-token ratio	TYPETOKN	Coh-Metrix	The number of unique types divided by tokens

controlling for multicollinearity, 17 variables were entered into a stepwise DFA and a stepwise LOOCV DFA.

The model created by the stepwise DFA achieved a classification accuracy of 65.8%, using five variables (*Motion prepositions*, *Type-token ratio*, *Spoken range content words*, *Personal pronouns*, and *Insight words*); see Table 6 for a detailed description of these variables) which is significantly higher ($df = 9$, $n = 240$, $\chi^2 = 277.725$, $p < .001$) than the baseline accuracy of 25%. The reported Kappa = .544, indicates moderate agreement between actual and predicted task type (Landis & Koch, 1977). The stepwise LOOCV DFA achieved a classification accuracy of 60.4%, suggesting that the predictor model is stable across the dataset. Table 7 includes descriptive and MANOVA statistics for the variables included in the final predictor model.

Table 8 includes the confusion matrix for the stepwise DFA. Accuracy figures for each task type are calculated by dividing the number of correctly classified texts by the number of original texts to be classified. Task type group membership predictions were accurate for 56.7% of Independent – Personal Preference task responses, 48% of Independent – Choice task responses, 70.0% of Integrated – Campus Situation task responses and 91.7% of Integrated – Academic Course task responses. The confusion matrix shows that responses to the Independent – Personal Preference tasks and responses to the Independent – Choice tasks were most often misclassified as each other. The confusion matrix also shows that misclassified responses to the Integrated – Campus Situation tasks were spread among the other three task types, and that few responses to the Integrated – Academic Course tasks were misclassified. Table 9 comprises the results of a post-hoc Sheffe's test, which indicates the statistical differences and similarities between each subtask with regard to each index selected by the stepwise DFA. This table shows that no statistical differences were observed between the independent subtasks. Additionally, this table indicates that statistical differences were found between each of the independent subtasks and the Integrated – Academic Course with regard to each of

Table 7. Means (standard deviations), *F* values, and effect sizes for the four task types.

Variables	Independent personal preference	Independent choice	Integrated campus situation	Integrated academic course	*F* (3, 236)	η^2 p
Motion prepositions	60.501 (25.878)	54.329 (28.627)	62.942 (25.194)	25.444 (13.114)	31.254	0.284
Type-token ratio	56.896 (7.466)	56.647 (6.839)	49.386 (5.891)	51.734 (5.233)	20.107	0.204
Spoken range content words	3.553 (0.092)	3.555 (0.123)	3.486 (0.110)	3.215 (0.137)	114.452	0.593
Personal pronouns	0.119 (0.043)	0.116 (0.037)	0.087 (0.035)	0.047 (0.043)	43.680	0.357
Insight words	0.022 (0.016)	0.026 (0.019)	0.031 (0.016)	0.014 (0.011)	12.191	0.134

Note: For all indices, $p < .001$.

Table 8. Confusion matrix for Analysis 2 DFA: Four task types.

Actual task type	Predicted as Independent Personal Preference	Predicted as Independent Choice	Predicted as Integrated Campus Situation	Predicted as Integrated Academic Course	Total
Independent (Personal Preference)	**34**	19	7	0	60
Independent (Choice)	23	**27**	9	1	60
Integrated (Campus Situation)	3	9	**42**	6	60
Integrated (Academic Course)	0	2	3	**55**	60
Accuracy	56.7%	45.0%	70.0%	91.7%	**65.8%**

Table 9. Differences between task types with regard to predictor variables in Study 2.

Variable	Highest mean value				Lowest mean value		
Motion prepositions	Independent Personal Preference	=	Integrated Campus Situation	=	Independent Choice	>	Integrated Academic Course
Type-token ratio	Independent Personal Preference	=	Independent Choice	>	Integrated Academic Course	=	Integrated Campus Situation
Spoken range content words	Independent Personal Preference	=	Independent Choice	>	Integrated Campus Situation	>	Integrated Academic Course
Personal pronouns	Independent Personal Preference	=	Independent Choice	>	Integrated Campus Situation	>	Integrated Academic Course
Insight words	Integrated Campus Situation	=	Independent Choice	>	Integrated Academic Course		
		>	Independent Personal Preference	=			

Note: "=" indicates no significant differences; ">" indicates difference is significant at $p < .05$.

the five indices, and that statistical differences were found between the two integrated subtask types for every index with the exception of type-token ratio.

Discussion

In this study we investigated differences lexical and cohesive characteristics of responses to TOEFL iBT speaking tasks using advanced natural language processing tools (i.e.,

Table 10. Lexical and cohesive characteristics of task types.

Variable	High	Mid	Low
Motion prepositions	Independent Tasks		Integrated Academic Course
	Integrated Campus Situation		
Type-token ratio	Independent Tasks		Integrated Academic Course
			Integrated Campus Situation
Spoken range content words	Independent Tasks	Integrated Campus Situation	Integrated Academic Course
Personal pronouns	Independent Tasks	Integrated Campus Situation	Integrated Academic Course
Insight words	Integrated Campus Situation	Independent Tasks	Integrated Academic Course

Coh-Metrix, LIWC, and TAALES). We first explored differences between responses to independent and integrated tasks following previous research (e.g., Biber and Gray, 2013), and then took a more fine-grained approach by examining the effect of language use domain on linguistic production. The results indicate that there are lexical and cohesive differences between the responses to independent and integrated speaking tasks investigated in this paper, which is in line with previous studies. However, the results also indicate that the target language use domain of a task affects the lexical and cohesive characteristics of responses. The two independent tasks (Personal Preference and Choice) elicited responses that were very similar with regard to lexical and cohesive characteristics. This finding suggests that TOEFL iBT speaking tasks may fit more appropriately into three categories (Independent, Integrated – Campus Situation, and Integrated – Academic Course) than two (independent and integrated). The independent responses differed from responses to the Integrated – Academic Course task such that most linguistic features demonstrated inverse relations. The Integrated – Campus Situation task responses, however, shared some lexical and cohesive characteristics with both independent and Integrated – Academic Source responses. Table 10 provides an outline of the lexical and cohesive characteristics of the task types. These characteristics are also described below.

Combined independent tasks

The results indicate that responses to independent tasks are characterized by the use of less sophisticated lexis, more personal pronouns, more motion prepositions, a variety of words, and words that can be used to express an opinion. Specifically, independent responses include words that occur in a large number of spoken contexts in the BNC. They also include a high number of personal pronouns (e.g., I, my), motion prepositions

(e.g., from, through, up), and insight words (e.g., think, know, consider). These characteristics of independent task responses are not particularly surprising given the nature of the tasks, which ask test-takers to provide their opinion on relatively broad topics. As one might expect, the data indicates that responding to these types of tasks does not require specialized or infrequent vocabulary or the repetition of ideas. Additionally, it is common for test-takers to use personal pronouns and insight words when responding to a task that asks them to express an opinion or defend a choice.

Integrated – Campus Situation tasks

The results indicate that responses to Integrated – Campus Situation tasks are characterized by the inclusion of less sophisticated lexis, the repetition of words, the use of personal pronouns, the use of motion prepositions, and include a high number of insight words that may be characteristic of providing an opinion. Specifically, these responses include words that occur in a large number of spoken contexts in the BNC. These responses also have a low type-token ratio, include a moderate number of personal pronouns (fewer than independent tasks but more than Integrated – Academic Course tasks) and motion prepositions, and include a high number of insight words. Again, the textual characteristics of the Integrated – Campus Situation tasks felicitously align with the nature of the tasks, which ask test-takers to summarize a (ostensibly common) problem that a student or professor is having and the solutions that are suggested by the interlocutor and then provide their own opinion on the appropriate solution to the problem. Much like the independent tasks, responding to Integrated – Academic Course task does not require particularly specialized lexis because the problem in question is common. Unlike the independent tasks, however, the Integrated – Academic Course task requires responses to include a summary of a problem and potential solutions given by the speakers, resulting in the repetition of words and ideas. Finally, much like the independent tasks, the Integrated – Academic Course task requires test-takers to provide an opinion regarding topic (in this case a solution to the problem), and accordingly a moderate number of personal pronouns and a high number of insight words (e.g., think, know, consider) are used.

Integrated – Academic Course tasks

The results indicate that responses to Integrated – Academic Course tasks are characterized by the use of sophisticated lexis, the repetition of words and ideas, and an avoidance of language related to the expression of personal opinion. Specifically, these tasks elicit responses that include words that occur less frequently in the spoken portion of the BNC. In addition, these responses tend to have low type-token ratios. The responses also include a low number of personal pronouns, a low number of insight words such as *think*, *know*, and *consider*, and a low number of motion prepositions. These textual characteristics seem to align with the task requirements, which ask test-takers to simply summarize the main points in a simulated academic lecture without providing their opinion of the topic. Given that the lectures are constrained to specialized topics, summaries of the lectures unsurprisingly include sophisticated lexis and repeated words and ideas (those that were presented in the lecture). Because test-takers are not asked to provide their opinion, they tend to avoid using personal pronouns and insight words. In addition, like the

Integrated–Campus Situation task, the Integrated–Academic Course requires responses to include a summary of a problem, resulting in the repetition of words and ideas.

Conclusion

This paper has explored the construct validity of the inclusion of multiple task types in the speaking section of the TOEFL iBT. Using powerful natural language processing tools, we identified three distinct task types, and used the textual, linguistic characteristics of responses to these task types to accurately classify them using discriminant function analyses. The findings of this project provide further evidence to support the TOEFL iBT validity argument, and specifically the explanation inference. The inclusion of tasks that vary with regard to source-use and language use domain appear to elicit a variety of lexical features in successful task responses. Furthermore, the findings of this study have important implications for future research.

The findings of this project have provided support for the inclusion of different task types for the speaking portion of the TOEFL iBT. The findings indicate that the two independent tasks elicit responses with similar linguistic features, which in turn are distinct from responses to Integrated – Academic Course tasks. Responses to Integrated – Campus Situation tasks are also distinct in some aspects, but share features with responses to both independent tasks and Integrated – Academic Course tasks. These findings partially support previous studies on TOEFL iBT speaking tasks, in that independent tasks share similar features and that integrated tasks have some shared features (e.g., Barkaoui et al., 2013; Biber & Gray, 2013; Brown et al., 2005). The findings diverge from previous studies, however, in that integrated responses were not found to elicit responses that were uniformly similar. Differences in language use domain (i.e., Campus Situation vs. Academic Course) affected the lexical and cohesive characteristics of responses to integrated tasks. The inclusion of a source-based task component results in responses that differ from independent tasks, but language domain is also an important factor for response characteristics and one that leads to different linguistic production on the part of test-takers.

As suggested by the findings of this project, future research into independent and integrated tasks should be sensitive to potential differences between task types, regardless of input type. Future research should also address the limitations of this project. First, we only investigated four of the six TOEFL iBT speaking tasks. Clearly, the linguistic features of the two read/listen/speak task types should be investigated with the caveat that read/listen/speak tasks may not wholly represent test-takers' linguistic abilities. Such investigation will bring to light the degree to which language use domain differences (such as Campus Situation and Academic Course) and type of input (listening versus listening and reading) affect the linguistic features of responses (and, perhaps, the relative importance of each of these factors). Second, future research should investigate features that we did not investigate in this study, for example, features of pronunciation, such as prosody. Third, this project has determined that there are differences between responses to certain tasks, and that these differences can be used to create accurate predictor models. We have not, however, tied these linguistic features to authentic, definable constructs by tying the linguistic features of each task type to the linguistic features of similar, authentic situations. Research is certainly warranted in this area.

Acknowledgements

We would like to add the following acknowledgements: "We would like to express our thanks to Educational Testing Service for providing the public use data set. We would also like to thank to Diane Belcher, Sara Weigle, and Sarah Goodwin for helpful feedback on earlier drafts of this paper.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics (Oxford)*, *34*(3), 304–324.

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

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis* (TOEFL iBT Research Report-19). Princeton, NJ: Educational Testing Service.

Brown, A., Iwashita, N., & McNamara, T. (2005). An Examination of Rater Orientations and Test-Taker Performance on English-For-Academic-Purposes Speaking Tasks. *ETS Research Report Series*, 2005(1), i–157.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, *40*(2), 540–545.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioural Research Methods Instrumentation and Computers*, *16*, 502–532.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3): 904–911.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45–78). Cresskill, NJ: Hampton.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language*. New York: Routledge.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*(2), 213–238.

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. Language Learning & Technology, 17, 171–192.

Crossley, S., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*(3), 250–270.

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

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238.

Hamp-Lyons, L. (1991). Issues and directions in assessing second language writing in academic contexts. In L. Hamp-Lyons (Ed.). *Assessing second language writing in academic contexts* (pp. 323–330). Norwood, NJ: Ablex.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*(2), 282–306.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods*, *44*(4), 978–990.

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly 49*(4), pp. 757–786. doi: 10.1002/tesq.194

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131–166.

McNamara, D.S., Graesser, A.C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*, vol. I: *Transcription format and programs* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Malvern, D. D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, UK: Palgrave Macmillan.

McNamara, D. S., Graesser, A. C., Cai, Z., & Kulikowich, J. M. (2011). Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension. In annual meeting of the American Educational Research Association, New Orleans, LA.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin, TX: LIWC.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for specific purposes*, *9*(2), 109–121.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics (Oxford)*, *31*(4), 487–512.

Svartik, J., & Quirk, R. (1980). *A Corpus of English Conversation*. Lund: Gleerup.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available from: www.natcorp.ox.ac.uk/.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Wang, L., Eignor, D., & Enright, M. K. (2007). A final analysis. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the TOEFL iBt* (pp. 259–318). New York: Routledge.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*(3), 335–353.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895.