Article

LANGUAGE
TESTING

Language Testing
1–23
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0265532219860750
journals.sagepub.com/home/ltj
SAGE

Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content

Kellie Frost
University of Melbourne, Australia

Josh Clothier
University of Melbourne, Australia

Annemiek Huisman
University of Melbourne, Australia

Gillian Wigglesworth
University of Melbourne, Australia

Abstract

Integrated speaking tasks requiring test takers to read and/or listen to stimulus texts and to incorporate their content into oral performances are now used in large-scale, high-stakes tests, including the TOEFL iBT. These tasks require test takers to identify, select, and combine relevant source text information to recognize key relationships between source text ideas, and to organize and transform information. Despite being central to evaluations of validity, relationships between stimulus content, task demands, and the oral discourse produced by test takers are yet to be empirically scrutinized to an adequate degree. In this study, we focus on a TOEFL iBT reading–listening–speaking task, applying discourse analytic measures developed by Frost, Elder and Wigglesworth (2012) to 120 oral performances to examine (a) the integration of source text ideas by test takers across three proficiency levels, and (b) the appropriateness of content-related criteria in the TOEFL integrated speaking rubric. We then combine analyses of these aspects of performances with a qualitative analysis of the generic structure and semantic profiles of stimulus texts to explore relationships between stimulus text properties and oral

Corresponding author:
Kellie Frost, School of Languages and Linguistics, University of Melbourne, Parkville, Victoria, Australia.
Email: kmfrost@unimelb.edu.au

performances. Findings suggest that the extent to which content-related rating scale criteria distinguish between proficiency levels is contingent on stimulus text properties, with important implications for construct definitions and task design.

Keywords
Discourse analysis, EAP assessment, integrated tasks, rating criteria, speaking assessment

Challenges to defining the construct of integrated speaking tasks

Language tests, including performance-based tests, have traditionally focused on measuring independent constructs of the four skills: speaking, listening, reading, and writing. In the context of university study, and other "real world" contexts, communicative acts rely on the integration of two or more of these skills, as well as other non-linguistic cognitive abilities (Douglas, 1997, 2000). To more closely mirror the demands faced by students entering English-dominant tertiary institutions, integrated tasks of speaking and writing are now used in large-scale, high-stakes tests, including the TOEFL iBT. Integrated test tasks are those that require test takers to listen to or read source texts, and then incorporate information from these texts into spoken or written responses (Lewkowicz, 1997).

Integrated English for academic purposes speaking and writing tasks involve complex texts requiring test takers to engage cognitive skills which extend beyond language proficiency skills. These cognitive skills include identifying, selecting, and combining relevant information from academic texts into oral and written performances, recognizing key relationships between source text ideas, and organizing and transforming relevant content (Brown, Iwashita, & McNamara, 2005). In targeting such skills, which are highly valued in tertiary education domains, integrated test tasks offer more authentic and comprehensive construct representations of academic speaking and writing ability. However, empirical research into relationships between stimulus text content, the demands made on test takers at different levels of proficiency, and the oral discourse produced by test takers, remains scarce. Consequently, how test takers engage with and make use of ideas from source texts in response to integrated speaking tasks remains, largely, intuited.

It is broadly agreed that the inclusion of integrated tasks in speaking and writing tests will yield more appropriate evidence of academic language proficiency than the use of independent tasks alone, and previous studies have highlighted several potential issues with implications for task design and rating scale development (Barkaoui, Brooks, Swain, & Lapkin, 2013; Brown et al., 2005; Crossley, Clevinger, & Kim, 2014; Cumming, Kantor, Baba, Eouanzoui et al., 2005; Cumming, Kantor, Baba, Erdosy et al., 2005; Frost et al., 2012; Lee, 2006; Plakans, 2009; Plakans & Gebril, 2012; Weigle & Parker, 2012). In relation to speaking assessment, Frost et al. (2012) investigated a prototype listening-to-speak task developed by Oxford University Press and found that although content-related aspects of oral performances distinguished between proficiency levels, test takers, regardless of proficiency, overwhelmingly reproduced source text information idea-unit for idea-unit, rather than paraphrasing or summarizing information. These findings raised questions about the appropriateness of including "input well summarized" as part of rating scale criteria, and about the extent to which such tasks tap into summarization processes considered relevant to academic domains. Questions have also been raised about the consistency with which raters distinguish between comprehension and production abilities in accounting for test takers' source text use in oral performances, especially in relation to content inaccuracies (Brown et al., 2005). Lee (2006) and Barkaoui et al. (2013) have also raised the possibility that different stimulus materials elicit different cognitive processes, thus potentially tapping into different constructs.

These issues, which highlight the importance of empirically examining relationships between the content-related characteristics of stimulus texts and test taker performances, have yet to be adequately addressed in existing research. To our knowledge, only Frost et al. (2012) and Crossley et al. (2014) have operationalized the relationship between stimulus content and the content produced by test takers in integrated speaking tasks. Frost et al. (2012) showed that content-related aspects of performance distinguished between proficiency levels, providing support for some content-related scoring criteria but offering little evidence of the use of summary skills. Crossley et al. (2014) examined the relationship between characteristics of stimulus listening materials and the incorporation of stimulus text words into oral performances by test takers in response to a TOEFL iBT listening–speaking integrated task. Crossley et al.'s (2014) study was based on the established theoretical premises that recall ability depends on word properties such as abstractness and concreteness (e.g. Paivio, 2007; Paivio, Yuille, & Madigan, 1968), and that recall ability underlies an ability to summarize information from source texts (e.g. Kintsch & van Dijk, 1978). The researchers examined whether word-level properties of listening stimulus materials influenced how easily words could be recalled and thus integrated by test takers into spoken responses. They also examined how the use of source text words by test takers predicted human judgments of their speaking ability. They found that word-level properties, including the incidence of word occurrence in the source text and the use of words in positive connective clauses, were strong predictors of word integration into oral performances, and that the incidence of source text words in spoken responses was a strong predictor of human judgments of speaking quality.

Although Crossley et al.'s (2014) study revealed important insights into the relationship among stimulus content, test taker performances, and test scores in relation to a TOEFL iBT integrated listening-to-speak task, their analysis remained at the level of individual words and clauses. Recent work in the sociology of education highlighted a need to examine characteristics of whole stimulus texts, rather than word- and clause-level properties, in order to gain insights into potential task demands (Freebody, 2013; Maton, 2013, 2014, 2016; Matruglio, Maton, & Martin, 2013). In integrated speaking tasks, this suggests that task complexity, especially the demands involved in integrating concepts and ideas from source texts into oral performances, is contingent on how these concepts and ideas are introduced, exemplified, and developed within and across stimulus materials.

Maton (2013, 2014, 2016) extended Paivio et al.'s (1968) notions of abstractness and concreteness, as utilized by Crossley et al. (2014), by situating the semantic properties of texts as dynamic and emergent, located along a continuum rather than existing as fixed, inherent properties of decontextualized concepts or words. According to Maton (2013), levels of abstractness and concreteness are relative and highly contextual, emerging and shifting across texts in academic contexts as concepts are introduced, explained, and exemplified. His heuristic approach on mapping the semantic profile of texts (i.e. the movements between levels of abstractness and concreteness that characterize the development of particular concepts or ideas within texts) has been adopted in studies investigating literacy development (Martin, 2013) and the relationship between teaching practices and knowledge building in schools (Freebody, 2013; Matruglio, Maton, & Martin, 2013). These findings suggest a relationship between the semantic profiles of both written and spoken input texts to which students are exposed, and the quality of texts students produce. Maton (2013) further argued that concept and idea development that follows semantic "wave" profiles (i.e. shifting back and forth between more abstract and more concrete manifestations of ideas) provides better scaffolding for comprehension, recall ability, academic skill development, and knowledge building. This is because such development models the attributes of "ideal" oral and written texts for students, since an ability to provide explanations that bridge the gap between more highly decontextualized concepts, and more local, concrete, and context-dependent ideas is highly valued in academic domains.

In this study, we move beyond word- or clause-level analyses of stimulus texts, towards examining how information is introduced, developed, and exemplified within and across academic texts. We do this by examining both the semantic properties and generic structure of stimulus texts. Since the latter is thought to provide scaffolding for test takers (Brown et al., 2005; Plakans & Gebril, 2012), both potentially affect test takers' ability to recall main ideas, and identify and articulate conceptual relationships between ideas in oral performances. Although findings from existing research into integrated speaking tasks highlight the importance of combining a content analysis of test takers' oral performances with an analysis of the properties of stimulus texts, further work is needed to generate empirically robust construct definitions, and evaluate the appropriateness of task design and rating scale criteria, particularly where rubrics make explicit reference to the use of content.

Study aims

In this study, we combine an analysis of the content-related dimensions of test takers' oral performance discourse in a TOEFL iBT integrated reading–listening-to-speaking task with a more holistic, qualitative analysis of the semantic profiles and generic structure of the task stimulus materials. The TOEFL iBT speaking section includes two integrated reading and listening-to-speak tasks, one based on campus life situations and one based on academic content. Given our aim to examine the development, within and across texts, of academic concepts associated with knowledge building, using Maton's (2013) notion of semantic profiles, we focused solely on the latter task in the current study. Study aims were as follows:

1. To investigate whether content-related aspects of task performances relevant to the TOEFL iBT integrated speaking rubric distinguished between test takers of different proficiency levels; and

2. To identify potential relationships between stimulus text properties and the characteristics of content-related aspects of test taker performances at different proficiency levels.

The reading–listening-to-speaking task enabled us to examine the potential relationships between the content produced by test takers and how concepts and ideas were developed across different stimulus text modes. To do this, we selected discourse analytic measures developed by Frost et al. (2012) that were consistent with the TOEFL iBT integrated speaking task scoring criteria[1] for performance content, which, under the heading "topic development," distinguish between levels based on accuracy of content, completeness of coverage of relevant information, and progression of ideas. We also incorporated the notion of semantic profiles (as developed by Maton, 2013) into an analysis of the generic structure of the stimulus materials to describe qualitatively how main ideas develop across the reading and listening stimuli.

Methods

The TOEFL integrated reading-listening-speaking task

The TOEFL iBT is a computer-delivered academic English test, designed to assess English language skills in readiness for studies in English-dominant universities. The speaking section consists of two independent tasks (questions 1 and 2), and four integrated tasks: two reading–listening–speaking tasks (questions 3 and 4) and two listening–speaking tasks (questions 5 and 6). Questions 3 and 5 are based on campus life situations; questions 4 and 6 on topics representative of academic course content.

Question 4, the focus of this study, requires test takers to read a short passage in 50 seconds, and then listen to a mini lecture of approximately one minute's duration on the same academic topic. The reading text introduces and generally explains the topic, and the listening text provides a specific example of the topic. While reading and listening to each text, test takers may take notes and after the mini lecture, 30 seconds preparation time is provided. Test takers are then presented with a prompt to explain a given aspect of the topic, using the example from the mini lecture. Speaking time is one minute.

We used two parallel versions of this task from speaking test forms provided to us by the Educational Testing Service (ETS). Form 1, "Allergies," related to the subject area of biology, and Form 2, "Sunk Costs," was in the domain of economics.

Dataset

The dataset consisted of stimulus materials for the reading–listening–speaking task from two test forms, including a reading text and audio recording and transcript of the mini lecture for each, and 120 audio recordings and transcripts of test taker oral performances across the two versions (60 for each), all provided to the researchers by ETS.

Stratified random sampling was used to select the 120 test taker performances from the 480 supplied to us by ETS to ensure a spread of scores and a balance across genders. TOEFL overall raw scores for the speaking section, which can range from 0 to 24, were also provided to us by ETS. Using these scores, 20 performances were allocated to each of three proficiency levels: Low (raw score range 14–17); Middle (range 18–19); High (range 20–24). Although we acknowledge that the overall score will be influenced by the task, and is thus to some extent circular, the aim was to ensure a spread of proficiency levels.

Data preparation

The reading and listening text transcripts and test taker performance transcripts were segmented into idea units. Idea units were defined in accordance with criteria provided by Frost et al. (2012), which had been adapted from an earlier definition provided by Kroll (1977) to include the following:

1. all clauses, including subordinate and relative clauses;
2. sub-clause variations are also considered as idea units, according to the following parameters:
    a. coordinated verb phrases are counted as separate idea units;
    b. phrases acting as discourse markers, typically set off from related clauses by commas, are considered to combine with related clauses as a single idea unit;
    c. coordinated nouns or noun phrases connected to a common verb phrase are counted as separate idea units;
    d. coordinated *independent* adjectives connected to a common verb phrase are counted as separate idea units; and
3. illustrating or clarifying examples are individual idea units, even where included in a clause.

Coding procedures

We coded the segmented test taker performance data using NVivo version 11, using select discourse analytic measures developed by Frost et al. (2012). The measures were selected to correspond with scoring criteria in the TOEFL iBT integrated speaking rubric, which specify the following: accuracy of content, the completeness of coverage of relevant information, and the progression of ideas, as set out below. We coded the stimulus materials for generic structure, based on approaches adopted by Brown et al. (2005) and Frost et al. (2012), and the semantic profiles of the text were qualitatively described according to Maton's (2013, 2014) framework (see the "Stimulus material analysis" section).

Test taker performance data

Accuracy of content

Test taker idea units were either *accurate* or *distorted* reproductions of corresponding source text idea units. Reproductions were accurate if they captured the same meaning presented in source text idea units, regardless of lexical or grammatical errors. These were unit-for-unit reproductions and instances where test takers accurately reduced two or more source text idea units into a single idea unit. For example:

> Allergies text idea units: *these dust mites contain // and release proteins*
>
> Test taker idea unit: *proteins was released by dust mites*
>
> Sunk Costs text idea unit: *you could be watching the same game at home*
>
> Test taker idea unit: *I can see the same football match in my home*

Distortions involved an inappropriate change in meaning, which sometimes stemmed from lexical or grammatical errors. For example:

> Allergies text units: *to protect itself//against invading substances*
>
> Test taker idea unit: *It's a kind of protection from harmonious substances*
>
> Sunk Costs text idea unit: *Sunk costs can affect people's decisions*
>
> Test taker idea unit: *people are always influenced by the decisions*

Completeness of coverage of relevant information

We operationalized coverage of relevant information as the number of stimuli main ideas reproduced in test taker responses. The main ideas for each task were derived by six staff members at the Language Testing Research Centre, University of Melbourne, who examined the stimulus reading and listening texts for both tasks under the same time pressure as applies in the test. Each staff member independently derived a list of main ideas and the researchers collated ideas over which there was consensus to produce a list of seven main ideas for each task, numbered from 1 to 7 in the order in which they first appeared in the stimulus materials. See Table 1 and Figure 2 (in the "Results" section), in which they are situated within the generic structure for each set of stimulus materials. Accurately reproduced main ideas in test taker performances were coded according to the corresponding main idea number from the input materials.

Progression of ideas

We operationalized test taker progression of ideas in the order in which main ideas occurred in test taker performances compared to where they occurred in the generic structure of the stimulus materials (see Table 6 and Figure 2).

Stimulus material analysis

Given the stimulus materials consisted of reading and listening texts, our analysis of the generic structure of the stimulus materials involved identifying obligatory and optional

Table 1. Summary of main ideas by topic.

	Allergies	Sunk Costs
Main idea 1	*Allergic reaction: immune system/body* ***reacts against harmless substances*** *(allergens)* ***mistakes them as threat***	*Sunk Cost =* ***money*** *that is invested* ***cannot be recovered*** *if a project is abandoned*
Main idea 2	*Immune response/fight against allergens →* ***allergy symptoms***	***people continue projects that should be discontinued****/act against own best interest (i.e. keep investing)*
Main idea 3	***particles in dust*** *(not dust) OR* ***dust mites*** *→ allergies*	*Keep investing* ***because of the money they have already spent***
Main idea 4	*Dust mites contain* ***proteins*** *that we breathe in*	***Spent money on football game ticket***
Main idea 5	*Immune system* ***makes antibodies****/ substances to fight "invaders"*	***Bad weather*** *on the night of the game*
Main idea 6	*Antibodies cause cells to* ***release chemicals***	*It's* ***not in self-interest to go to the game*** *(you don't want to go, you're better off staying home, warm & cozy)*
Main idea 7	*Chemicals irritate nose, eyes, throat/****Chemicals cause symptoms***	*There is an alternative option:* ***same game is on TV***

generic "stages" (see Swales, 1990, and Derewianka, 2003, for written texts; Eggins & Slade, 1997, for oral texts). The reading texts for both task topics followed an "explanation" structure (Derewianka, 2003) containing a statement of a phenomenon to be explained, an "introduction" stage, and an "explanation" stage in which details of the phenomenon were provided. The listening texts for both topics mirrored the generic structure of an anecdote: an account of a remarkable event or problematic experience (Eggins & Slade, 1997). As such, the structure included obligatory "complication" and "resolution" stages, through which the event or experience is conveyed. In addition, there were three optional stages: an "abstract" stage to signal that a story is about to be told, an "orientation" stage to provide details of people, time, and place, and an evaluation stage involving appraisals of the event or experience as the story unfolds (Eggins & Slade, 1997). An outline of the generic structure of the stimulus materials for each topic is provided in Table 6 and Figure 2.

The semantic profile analysis of the stimulus materials for each topic involved identifying and mapping relationships between stages (and the main ideas expressed within stages) across the reading and listening texts for each topic. As noted above, a semantic profile (Maton, 2013) consists of the movements between levels of abstractness and concreteness that characterize the development of particular concepts or ideas within texts. Maton (2013) preferred the term "semantic gravity," defined as "the degree to which meaning relates to its context" (p. 11), over the terms "abstract" and "concrete," and we adopt his terminology in our analysis. Increasing semantic gravity involves a move within a text from "abstract or generalized ideas towards concrete and delimited cases" (2013, p. 11), whereas decreasing semantic gravity involves a move from specific examples towards more abstract or low context-dependent concepts.

Table 2. Codes for shifts in semantic gravity.

Semantic gravity (SG)	Code	Definition
Lowest SG (most de-contextualized)	Abstraction	Presents a general principle/phenomenon
	Summarization (Abstract)	Summarizes information already presented about general principle/phenomenon, including re-wording and restructuring of information. Does not present new information.
	Generalization	Presents a general observation or draws a general conclusion about the phenomenon *in relation to a category (e.g. allergy sufferers)*
	Interpretation	Relates information from illustrative case back to general principle/phenomenon
Highest SG (most contextualized, highest level of specificity)	Exemplification	Describes a *specific, illustrative case* of general principle/phenomenon

In order to identify shifts in semantic gravity across the main ideas expressed in successive stages across reading and listening texts for both topics, we adapted and applied the coding scheme or "translation device" presented by Maton (2014, p. 113) to the stimulus texts used in our study (see Table 2).

As shown in the far-right column of Table 2, we adopted a relative view of semantic gravity based on Maton's (2013, 2014) conceptualization of concreteness and abstractness as dynamic and contingent properties, which shift as ideas develop throughout the text(s). We illustrate this with main idea 2 in Allergies. The notion is that allergy symptoms occur because the immune system reacts mistakenly and fights a harmless substance. This main idea appears three times in the stimulus materials: at the end of the reading text, and the beginning and end of the listening text. The segment of the reading text and the first segment of the listening text corresponding to main idea 2 are compared below:

> Reading text segment: *The unpleasant symptoms that an individual with allergies experiences all result from the body's attempt to fight off a non-existent threat.*
>
> Listening text segment: *Well, there wasn't a day that went by without Joe having a runny nose, or watery eyes, and he just couldn't stop sneezing. One day Joe told me that the sneezing and all the other stuff was the result of him being oversensitive to dust.*

In the reading text segment, in terms of context-dependency, main idea 2 is manifest in a generalized case (*an individual with allergies*), and in non-specific terms that can take on different meanings across a range of contexts (*unpleasant symptoms, a non-existent threat*). Referring to Table 2, above, this manifestation of main idea 2 was coded as "Generalization," as it generalizes the more abstract notion of immune systems reacting mistakenly (main idea 1) to the case of individuals with allergies. In the listening text, the same main idea is presented via an increase in semantic gravity to “Exemplification” (see Table 2); that is, a shift to concrete and context-specific terms, detailing the experiences of a particular allergy sufferer, Joe. For example, *unpleasant symptoms* becomes a list of specific symptoms that Joe regularly experienced: *runny nose, watery eyes, sneezing. An individual with allergies* becomes *Joe* and the trigger of the allergy symptoms, *a non-existent threat* becomes specific and concrete in Joe’s case: *dust*.

We traced increases and decreases in semantic gravity across main ideas, from generic stage to generic stage as the texts unfolded, to produce a graphic representation of the semantic profile of the stimulus texts for each topic (see Figures 2 and 3). We then examined whether differences in the semantic properties of main ideas within stages were associated with the reproduction of main ideas by test takers. We did this by examining the semantic profiles of source text main ideas that test takers, regardless of proficiency, were able to reproduce with those that only high-proficiency test takers could reproduce.

Inter-coder reliability

Two researchers initially coded the entire dataset independently, and then discussed and resolved differences, and decided on final codes. A third researcher independently coded 20% of the data for the purpose of verifying inter-coder reliability, which was determined to be 95% for idea unit coding and 85% for main ideas and generic structure.

Statistical analysis

To examine differences between groups in the accuracy of content, we fitted mixed-effects logistic regressions (generalized linear mixed-effects regression (glmer), generalized linear mixed-effects models (glmm)) to model the variability in the data using the glmer function in the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R (RCore Team, 2017), RStudio IDE (RStudio Team, 2016). We chose this approach in the first place because ANOVA is not strictly compatible with outcome variables in this study which are frequency counts of categorical data; a logistic regression is better suited to this kind of data (Jaeger, 2008; Linck & Cunnings, 2015) as the incompatibility between ANOVA and categorical data “can lead to spurious null results and spurious significances” (Jaeger, 2008, p .435). Second, we sought to account for the effects of each test taker’s contribution to the overall variance in the model without our interpretation of between-group differences being inflated by any extraordinary performance by exceptional individuals in either direction. Mixed-effects modelling allows us to limit the effect of the clustering of each individual’s responses on the group contribution to the model; that is, the model can control for the clustering of variance on a per-subject (test taker) basis (the random effect) as well as identifying the statistical significance of the fixed effects of interest (Agresti, 2007, pp. 297–298; Cunnings, 2012; Link & Cunnings, 2015). This provides a reliable means of modelling count or categorical data with multiple observations per test taker. In our study, models were specified with random intercepts fitted for each test taker, and model fit was evaluated using likelihood ratio tests. We describe each model and the outcome of model fitting in the “Results” section. Post-hoc analysis was performed by comparing estimated marginal means using the emmeans

Table 3. Summary of idea unit production by proficiency group.

Allergies	Low		Middle		High	
	Number	Proportion[a]	Number	Proportion	Number	Proportion
Total idea units[b]	205		253		310	
Accurate idea units	137	66.8%	202	79.8%	245	79.0%
Distorted idea units	68	33.2%	51	20.2%	65	21.0%
Sunk Costs						
Total idea units	239		285		317	
Accurate idea units	187	78.2%	235	82.5%	276	87.1%
Distorted idea units	52	21.8%	50	17.5%	41	12.9%

[a] Proportion is reported as the percentage of the total idea units per proficiency group to adjust for differences in the lengths of samples.
[b] Units not specifically related to the input materials, such as invented ideas, were excluded from the analysis.

package in R (R Core Team, 2017), RStudio IDE (RStudio Team, 2016). For the reproduction of main ideas, glmer could not be fit because of the structure of the data. Glmer requires at least binary outcome variables, and this measure, when structured as binary data requires analysis of responses for each of the seven main ideas per task in turn, diluting the power of the analysis which requires larger Ns: the test would not be robust for these measures. Thus, we ran a Chi-square test of independence (a non-parametric test more robust to data that are distributed in this way) and associated *z*-tests for differences of proportions between cells (with Bonferroni adjusted *p*-values), testing for differences between number of main ideas produced by each proficiency group on a per main idea basis. The results of these tests are reported below.

Results

Accuracy of content

Table 3 summarizes the findings related to the accuracy of idea unit reproduction by proficiency group.

The overall number of idea units (amount of content reproduced) and the total number of accurately reproduced idea units increased with proficiency across both tasks, as did the proportion of total idea units in Sunk Costs. However, in Allergies, the high group produced more accurate and distorted idea units than the middle group; both middle and high groups produced a higher proportion of accurate idea units than the low group (66.8%). Test takers in all proficiency groups produced more idea units in response to Sunk Costs than Allergies, despite equivalent numbers of idea units across both tasks. Moreover, the low- and high-proficiency groups produced a higher proportion of accurate idea units for Sunk Costs than for Allergies (low group 78.2% vs. 66.8%, and high group 87.1% vs. 79%). For the middle group, the proportion was similar (82.5% and 79.8%, respectively). This is illustrated in Figure 1.

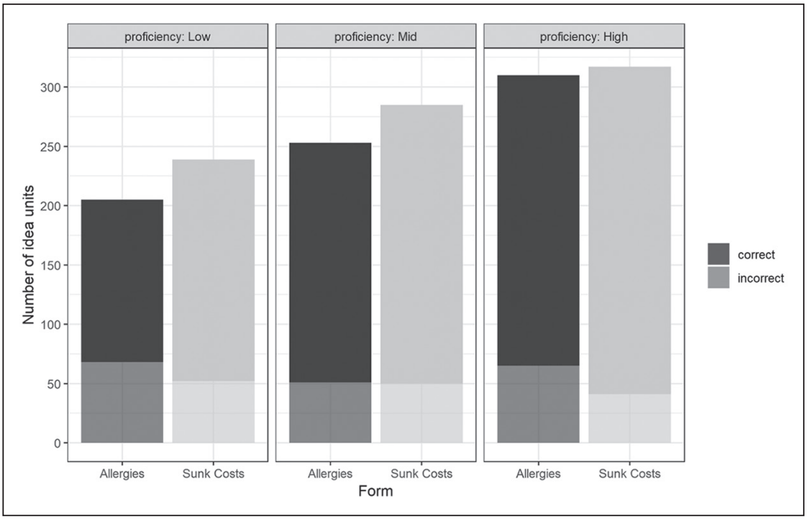


Figure 1. Accuracy of idea unit reproduction by proficiency and topic

To examine statistical significance, we fitted a mixed-effects logistic regression, with test taker specified as the random intercept term, proficiency (low, medium, high) as the predictor variable, and proportion of accurately (vs. inaccurately) reproduced idea units as the outcome variable. Performance of the model was tested using likelihood ratio tests. The best performing model had fixed main effects of proficiency and topic, and random intercepts for test taker nested within topic, χ^2 (1) = 4.5764, p =.03241. Averaged across the two topics, there was a significant difference between high and low groups, z = 2.816, p = .0135 (but not between middle and low groups, z = 2.335, p = .0513, or high and mid groups, z = 0.442, p = .8974). There was also a significant effect of topic: averaged across proficiency levels, participants completing Sunk Costs were less likely to supply incorrect idea unit reproductions than those completing Allergies, z = 2.438, p = .0148.

Data were also modelled separately for each topic. For Allergies, our model performed significantly better than a null model, with only the intercept and random terms specified, χ^2 (2) = 7.31, p = .02586. Pairwise comparisons between proficiency levels using Tukey's HSD show that differences observed in Table 3 and Figure 1 were significant when modelled using logistic regression between low and middle groups, z = 2.473, p = .0356, but not between other groups. Differences between the high and low groups approached significance (z = 2.327, p = .0521) due to individual differences in the high group; 3/20 high-group participants produced a high proportion of distortions in Allergies (35%, 38% and 54%). For Sunk Costs, proficiency provided no increased explanatory power over a null model (specified for only the random intercept term, test taker, of the model), χ^2 (2) = 2.8221, p = .2439, as there was less difference between groups on this topic in total idea units and proportion of accurately reproduced units compared to Allergies.

Table 4. Summary of reproduction of stimuli main ideas by proficiency.

Allergies	Low (n = 20)	Middle (n = 20)	High (n = 20)
Range	0–5	0–6	1–7
Mean no. of main ideas	2.45	3.95	5.1
SD	1.10	1.61	1.80
Sunk Costs			
Range	0–6	1–7	1–7
Mean no. of main ideas	3.55	4.45	5.2
SD	1.93	1.76	1.58

Table 5. Number of test takers by group producing individual main ideas.[a]

Allergies	MI1	MI2	MI3	MI4	MI5	MI6	MI7
Low	10	15	8	9	5	3	2
Middle	15	16	12	11	11	7	7
High	13	18	13	14	16	14	14
Sunk Costs	MI1	MI2	MI3	MI4	MI5	MI6	MI7
Low	10	10	10	12	11	11	9
Middle	9	15	16	19	11	12	7
High	14	14	14	18	17	14	14

[a] n = 20 in each group.

Completeness of coverage of relevant information

As discussed, seven main ideas were identified in the stimulus materials for each topic; Table 4 gives the mean number and range of main ideas reproduced by each proficiency group for both forms.

The mean number of accurately reproduced main ideas increased with proficiency for both topics, although the differences were not statistically significant as measured using Pearson's Chi-square test of independence, for either Allergies, χ^2 (12) = 10.914, p = .536, or Sunk Costs, χ^2 (12) = 4.479, p = .973. The large range of values resulted from the presence of individual differences within groups across both tasks (see the Appendix for further details).

We also examined the reproduction of each main idea within each proficiency group for each topic as shown in Table 5.

In Allergies, although more middle-proficiency test takers produced main idea 1 than high-group test takers did, for the remaining main ideas, numbers increased with proficiency. Main idea 2 in Allergies was reproduced most frequently, and accurately reproduced by over half of the test takers in all groups. Main idea 7 was only reproduced by two out of 20 low test takers, seven out of 20 in the middle group, but 14 out of 20 in the high group. We discuss the differences between main ideas 2 and 7 in Allergies under the heading *Generic structure and semantic profile*.

Table 6. Overview of generic structure: Allergies and Sunk Costs.

Stimulus	Allergies stages	Sunk Costs stages
Reading	Statement of phenomenon to be explained	Statement of phenomenon to be explained
	Explanation of phenomenon	Explanation of phenomenon
Listening	Orientation	Abstract
	Abstract	Orientation
	Micro-explanation of phenomenon	Complication (1a)
	Micro-explanation of phenomenon	Complication (1b)
	Complication	Complication (2a)
	Micro-explanation of phenomenon	Complication (2b)
	Micro-explanation of phenomenon	Evaluation
	Micro-explanation of phenomenon	Resolution
	Resolution	

Between-group differences were less evident in Sunk Costs. At least half the test takers in the low and middle groups accurately reproduced five out of seven main ideas, and 14 out of 20 in the high group reproduced all seven main ideas (see the Appendix). As Table 5 shows, for Sunk Costs, except for main idea 6, the number of main ideas produced did not increase consistently by proficiency, partly due to generic structure and semantic profile differences across the two sets of stimulus materials, discussed below.

Progression of ideas

To examine the progression of ideas by test takers in each proficiency group, we traced the order in which main ideas occurred in performances in each proficiency group compared to where main ideas were located in the generic structures of the stimulus materials (see Table 6 and Figure 2). For Allergies, all test takers closely followed the order of main ideas within the generic structure of the stimulus materials. In this study, 15 out of 20 test takers in the high-proficiency group followed the exact order, as did 17 in the middle-proficiency group, and 16 in the low-proficiency group. There were more deviations from the stimulus text structure for Sunk Costs. Half of the low-proficiency group changed the order in which main ideas were presented. Similarly, most of the middle group (12 out of 20 test takers) and most of the high group (11 out of 20 test takers) produced main ideas in a different order compared to the stimulus text structure. We discuss below the attribution of this finding to the different generic structures of the input materials for Allergies compared to Sunk Costs.

Stimulus material: Schematic structure and semantic profile

Table 6 provides an overview of the generic stages in the stimulus materials for Allergies and Sunk Costs. Figure 2 presents a more detailed view, including interrelationships between the reading and listening text stages, and an indication of where the seven main ideas for each topic were situated within stages.

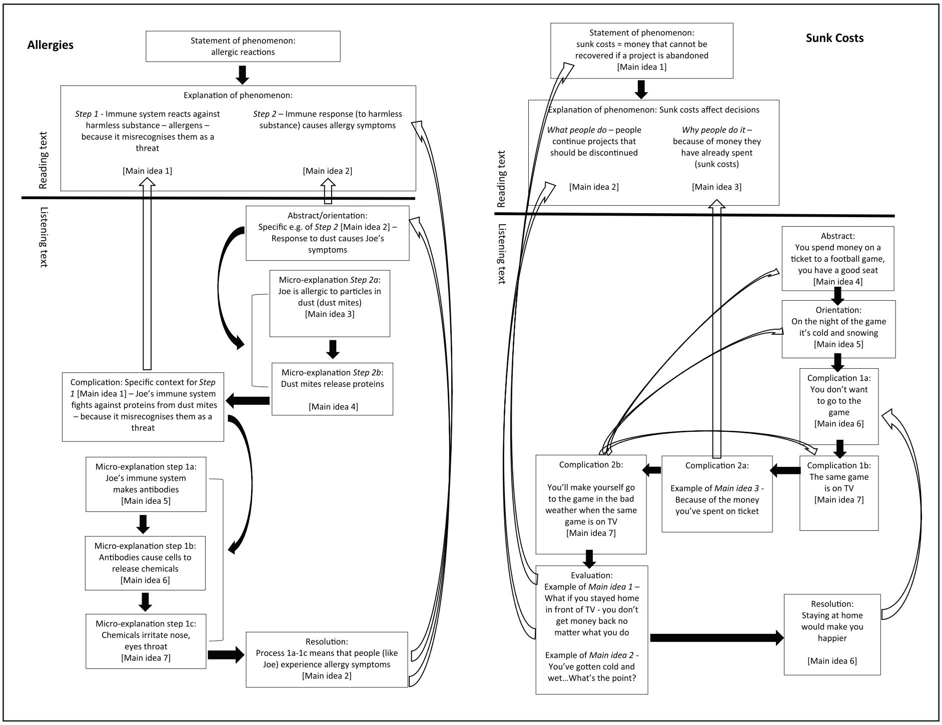

Allergies
Reading text
Listening text
Statement of phenomenon: allergic reactions
Explanation of phenomenon:
Step 1 - Immune system reacts against harmless substance – allergens – because it misrecognises them as a threat
[Main idea 1]
Step 2 – Immune response (to harmless substance) causes allergy symptoms
[Main idea 2]
Abstract/orientation: Specific e.g. of Step 2 [Main idea 2] – Response to dust causes Joe's symptoms
Micro-explanation Step 2a: Joe is allergic to particles in dust (dust mites) [Main idea 3]
Micro-explanation Step 2b: Dust mites release proteins [Main idea 4]
Complication: Specific context for Step 1 [Main idea 1] – Joe's immune system fights against proteins from dust mites – because it misrecognises them as a threat
Micro-explanation step 1a: Joe's immune system makes antibodies [Main idea 5]
Micro-explanation step 1b: Antibodies cause cells to release chemicals [Main idea 6]
Micro-explanation step 1c: Chemicals irritate nose, eyes throat [Main idea 7]
Resolution: Process 1a-1c means that people (like Joe) experience allergy symptoms [Main idea 2]
Sunk Costs
Reading text
Listening text
Statement of phenomenon: sunk costs = money that cannot be recovered if a project is abandoned [Main idea 1]
Explanation of phenomenon: Sunk costs affect decisions
What people do – people continue projects that should be discontinued
[Main idea 2]
Why people do it – because of money they have already spent (sunk costs)
[Main idea 3]
Abstract: You spend money on a ticket to a football game, you have a good seat [Main idea 4]
Orientation: On the night of the game it's cold and snowing [Main idea 5]
Complication 1a: You don't want to go to the game [Main idea 6]
Complication 1b: The same game is on TV [Main idea 7]
Complication 2a: Example of Main idea 3 - Because of the money you've spent on ticket
Complication 2b: You'll make yourself go to the game in the bad weather when the same game is on TV [Main idea 7]
Evaluation: Example of Main idea 1 – What if you stayed home in front of TV - you don't get money back no matter what you do
Example of Main idea 2 - You've gotten cold and wet…What's the point?
Resolution: Staying at home would make you happier [Main idea 6]

Figure 2. Allergies and Sunk Costs generic structure. (Black filled arrows indicate the progression of the texts from stage to stage, unfilled arrows with a solid line indicate explicit relationships between stages, which share main ideas and common lexical items as linking devices between segments of texts, and unfilled arrows with a broken line indicate implicit relationships between stages, that is, common main ideas, but an absence of explicit semantic links.)

Figure 2 shows for both topics, the reading-text main ideas recurred in the listening texts. As noted, in this task type, the reading text introduces and explains the phenomenon in question at a general level, and the listening text provides a specific example. While the stimulus materials for both topics followed the same overarching generic structure (explanation + anecdote), the Allergies listening text also contained an embedded "explanation" structure within the overarching anecdote structure. The embedded "explanation" structure was situated between the stages of "abstract" and "complication," and between "complication" and "resolution" (see also Table 6). The explanation stage encompassed a step-by-step micro-level explanation of how allergies occur.

Although the listening texts for both topics followed the generic structure of an anecdote, Allergies contained an embedded step-by-step micro-explanation of the biological processes that characterize the immune response that triggers allergy symptoms, meaning the order of main ideas needed to be maintained to make sense in the reproductions.

Differences in the generic structure and semantic profile of the stimulus materials potentially explain differences across the two tasks in the reproduction of main ideas.

Table 5 shows that for Sunk Costs, at least half the low group test takers accurately reproduced main ideas 1 to 6, with just under half accurately producing main idea 7. By comparison, only two out of seven main ideas in Allergies were reproduced by at least half of the low-proficiency group.

As illustrated in Figure 2, Allergies had three explicit links between the reading and listening texts related to main ideas 1 and 2. These were the only main ideas that were reproduced by at least half of all test takers, regardless of proficiency. In Table 5, main idea 2 was reproduced by the highest number of test takers in all groups (15/20, 16/20 and 18/20 in the low, middle and high groups respectively). Figure 2 shows that main idea 2 occurs across three stages, the *explanation* stage of the reading text (*step 2*) and the *abstract/orientation* stage and the *resolution* stage of the listening text. Thus, the high frequency with which main idea 2 was reproduced across all groups may be due to its high salience, enhanced by its semantic profile. The semantic profile of main idea 2, as discussed below, involves concept repetition, that is, the same idea manifest in different ways across reading and listening texts, with semantic "wave" movements (Maton, 2013) (shifts back and forth between abstract and concrete representations) in the listening text.

The Allergies reading text *explanation* stage provides a generalized, conceptual explanation of the immune system reaction that triggers allergies, and in the listening text *abstract/orientation* stage this explanation was tied to the specific case of Joe's specific symptoms (*runny nose, watery eyes, sneezing*), representing an increase in semantic gravity across these texts and stages (a shift from abstract to more concrete specific examples). In the listening text *resolution* stage, there is a shift from the specific case of Joe's allergy symptoms to allergy sufferers more broadly (*people like Joe*) and a more generalized process (*allergic reaction*), representing a decrease in semantic gravity. This serves to provide a link in the opposite direction, from the specific to the more conceptual explanation of the immune system and its role in generating allergic reactions in the reading text. These movements in semantic gravity are illustrated in Figure 3. Maton (2013) suggested this back and forth "wave" profile potentially provides scaffolding to facilitate comprehension and assist integration of ideas in production as important links between sections of text, and between concepts and specific examples, are reinforced and made more salient.

Comparing main ideas 2 and 7, the least frequently produced main idea, provides further support for the influence of the generic structure and semantic profile of the stimulus materials. Main idea 7 was reproduced by very few low- and middle-group participants: two and seven out of 20, respectively, compared to 14 out of 20 high-group participants. Main idea 7 occurs only once, towards the end of the listening text and represents the third and final step in the explanation of the micro-level biological processes underlying the occurrence of allergic reactions.

Main idea 7, expressed in the text by "these chemicals are what irritate the eyes, nose and throat," is the culmination of a specific process detailed across the three embedded micro-explanation stages (corresponding to main ideas 5, 6 and 7, respectively), and represents a series of increases in semantic gravity, beginning with the generalized concepts "immune system" and "allergic reaction" introduced in the explanation stage of the reading text. Across the micro-explanation steps 1a, b, and c in the listening text, the concept *immune system* is reduced to a specific aspect of the immune system (*antibodies*:

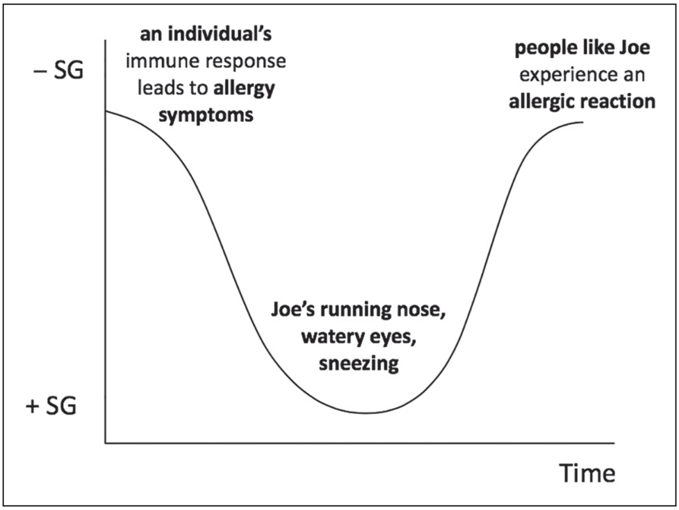


Figure 3. Semantic profile of Allergies main idea 2. In the figure, − SG indicates relatively low, and + SG relatively high, semantic gravity.

main idea 5), and the process *allergic reaction* is broken down into two micro-level processes: first, *antibodies cause cells to release chemicals* (main idea 6), and second, *chemicals irritate the eyes, nose, and throat* (main idea 7). Although there is repetition of key lexical items across these main ideas, there are no explicit semantic links between the specific micro-level aspects of the process encompassed by main ideas 5 to 7 and the more generalized, conceptual level explanation of allergic reactions in the reading text (see Figure 2). Furthermore, these main ideas were expressed only once, thereby limiting their salience.

The stimulus text structure in Allergies thus did not provide the same sort of explicit scaffolding between the general and specific for main idea 7, nor the repetition that occurred for main idea 2. The one-way downward movement towards increasing semantic gravity across main ideas 5 to 7 may hinder low-proficiency test takers' capacity to situate the specific step-by-step process in relation to more generalized processes outlined in the reading. An additional implication is that if test takers were unable to understand or recall any of main ideas 5 to 7, sense-making could only be maintained by integrating the conceptual level explanation provided in the reading into the micro-level process. This increased the cognitive complexity of Allergies compared to Sunk Costs, where no such integration was necessary.

In Sunk Costs (see Figure 2), in addition to implicit links between main ideas 1, 2, and 3 across reading and listening texts, there were several explicit links between main ideas 4, 5, 6 and 7, which may have increased the salience of relationships between main ideas within the listening text. Although the main ideas in the reading texts were highly generalized in both tasks, in the Sunk Costs listening text, all main ideas, including 1, 2, and 3, were predominantly manifest in highly specific, concrete representations (e.g. spending money on tickets to a football game) or particular familiar and everyday sensations and desires (e.g. feeling cold, wanting to stay home and watch TV). Moreover, there were no explicit links between the specific example in the listening text and the generalized explanation in the reading text, meaning low-level test takers could remain at the specific level in their oral performances, potentially reproducing main ideas without understanding or establishing links with the generalized concepts that the ideas exemplified; therefore, test takers did not need to integrate and combine information from both texts to make sense, as was the case in Allergies, where understanding and reproducing the main ideas in the embedded micro-explanation stages in the listening text required the integration of specific details and generalized concepts. Where scaffolding was absent, as in Allergies main ideas 5 to 7, most low-level test takers failed to reproduce them.

Discussion

Integrated speaking test tasks, which require test takers to integrate information from reading and listening stimulus texts into their oral performances, place cognitive demands on test takers that extend beyond the cognitive demands formerly associated with bare-prompt language proficiency tests (Douglas, 1997). Test takers need to identify relevant ideas in the sources, consider their interrelationships, and transform these mental connections into their oral texts (Brown et al., 2005). As discussed above, studies of integrated writing and integrated speaking tasks have suggested that characteristics of the stimulus texts may impact task complexity (Barkaoui et al., 2013; Lee, 2006; Plakans, 2009; Plakans & Gebril, 2012), the written or oral discourse produced (Crossley et al., 2014; Cumming, Kantor, Baba, Eouanzoui et al., 2005; Cumming, Kantor, Baba, Erdosy et al., 2005; Frost et al., 2012; Weigle & Parker, 2012), and rater judgments of performance quality (Brown et al., 2005; Crossley et al., 2014). Despite being central to evaluations of test validity, the content-related aspects of integrated speaking task performances, and their relationship to stimulus material characteristics and test score outcomes, are yet to be empirically scrutinized to an adequate degree. Such scrutiny is much needed to generate empirically and theoretically robust construct definitions and rating scale criteria. We contributed to this by comparing content-related aspects of test takers' oral performances in response to two parallel TOEFL iBT integrated reading and listening-to-speak tasks with a qualitative analysis of the task stimulus materials.

Consistent with Frost et al. (2012), the accuracy with which source text ideas were reproduced varied according to proficiency level. High-proficiency test takers reproduced more accurate source text ideas in terms of meaning, across both tasks, Allergies and Sunk Costs, than middle- and low-proficiency test takers. However, although for Allergies, between-group differences were significant or near significant, Sunk Costs was less effective in distinguishing between test takers. On the latter topic, all test takers, regardless of proficiency, produced a higher number and proportion of accurate idea units compared to Allergies.

High-proficiency test takers, on average, reproduced more stimulus text main ideas than low- and middle-group participants across both tasks, although inconsistencies across the tasks require some qualification. First of all, for Allergies, test takers in all proficiency groups closely followed the generic structure of the stimulus materials, whereas for Sunk Costs, there was substantial variation in the organization of information, regardless of proficiency. We attribute this to the different generic structures of the input materials for Allergies compared to Sunk Costs. Whereas the listening texts for both topics followed the structure of an anecdote, the Allergies listening text contained an embedded step-by-step micro-explanation of the biological processes that characterize the immune response involved in triggering allergy symptoms. The linear nature of this explanation meant that the order of stages needed to be maintained for test taker reproductions to make sense.

More notably, the number of main ideas produced by the low-proficiency group varied across tasks with six out of seven Sunk Costs main ideas accurately captured by at least half the participants, including those in the low-proficiency group. For Allergies, five of seven of the main ideas in the source texts were reproduced by less than half of the low-proficiency group members. We attributed this task-based difference to generic structure and semantic profile differences across the two sets of stimulus materials. In Sunk Costs, there were several explicit links between main ideas within the listening text, which likely increased their salience. Moreover, in the Sunk Costs listening text, all main ideas were manifest in highly specific, concrete representations, which enhanced the likelihood of test takers' integration of source text content. The absence of explicit links between the more abstract, conceptual explanation provided in the reading text meant test takers were not required to shift between the specific and conceptual in their performances in order to reproduce particular main ideas. In Allergies, by contrast, the reproduction of main ideas embodied in the micro-explanation stages in the listening text placed additional cognitive demands on test takers, requiring them to integrate specific details and generalized concepts, which required detection and understanding of implicit relationships between ideas across the texts. Where scaffolding was absent and the development of ideas was characterized by one-directional shifts in semantic gravity, as in Allergies main ideas 5 to 7, low-level test takers could not incorporate these ideas into their performances.

Our findings support Maton's (2013) argument that semantic waves serve to scaffold comprehension, bridging the gap between abstract, generalized ideas and more concrete and highly contextualized meanings. Allergies required test takers to provide explanations that bridge this gap in order to capture the main points from the input materials, whereas this was not the case in Sunk Costs. Our findings raise questions about the generalizability of integrated speaking tasks, consistent with concerns raised by Lee (2006), who suggested that different source texts may place different demands on test takers, and that supposedly parallel tasks may tap into different constructs.

Our study demonstrates a relationship between stimulus text characteristics, task demands, and the content-related aspects of test taker performances on integrated speaking tasks, which has important implications for task design and rating scale development. Although our findings provide some support for the content-related scoring criteria in the TOEFL iBT speaking rubric, they also raise questions about the generalizability of task performances resulting from different sets of stimulus materials. Task complexity appeared to be higher for Allergies than Sunk Costs, and as a consequence, Sunk Costs did not distinguish between test takers as effectively. Allergies, in which information in the stimulus materials followed a semantic wave profile, more closely mirrored an ideal academic text structure, making iterative links between generalized and decontextualized concepts and more concrete, specific examples. This type of generic structure and semantic profile is more likely effectively to elicit the skills involved in (a) selecting relevant information, and (b) identifying relationships between different ideas in source texts, skills highly valued in tertiary education settings. Incorporating these characteristics into task-design specifications could enhance the authenticity of tasks, and more effectively tap into content-related aspects of performance relevant to existing TOEFL integrated speaking task rating scale criteria.

Although the current study contributes to furthering understanding of relationships between input materials and the content-based aspects of test taker performances, which previous research indicates influences rater judgments and test outcomes (Brown et al., 2005; Crossley et al., 2014; Lee, 2006), the small sample size and a focus on performances across just two parallel tasks limits the conclusions that can be drawn. Differences in findings across the two tasks may have been due to the familiarity of the topics: Sunk Costs involved a phenomenon arguably more likely to be familiar to test takers (spending money for no gain), whereas Allergies involved technical terms and processes that may have been more difficult for those outside the domain of biology. Further research is needed to properly interrogate the impact of stimulus text characteristics on the comprehension and integration of source text information by test takers in response to integrated speaking tasks, with a combined focus on text properties, the content produced by test takers in speaking performances, and the cognitive processes and strategies engaged by test takers as they interact with these types of tasks.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by a TOEFL Committee of Examiners research grant.

Note

1. www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

ORCID iDs

Kellie Frost https://orcid.org/0000-0002-3424-3161
Josh Clothier https://orcid.org/0000-0001-9719-5585

References

Agresti, A. (2007). *An introduction to categorical data analysis*. New York: John Wiley & Sons. doi:10.1002/0470114754

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics*, *34*, 304–324. doi:10.1093/applin/ams046

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series MS-29). Princeton, NJ: Educational Testing Service. Available at www.ets.org/Media/Research/pdf/RR-05-05.pdf

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*, 250–270. doi:10.1080/15434303.2014.926905

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 1–75. doi:10.1016/j.asw.2005.02.001

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2005). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph Series MS-30). Princeton, NJ: Educational Testing Service. Available at www.ets.org/research/policy_research_reports/publications/report/2005/hsjg

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*(3), 369–382. doi:10.1177/0267658312443651

Derewianka, B. (2003). Trends and issues in genre-based approaches. *RELC*, *34*, 133–154. doi:10.1177/003368820303400202

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations*. (TOEFL Monograph Series MS-8). Princeton, NJ: Educational Testing Service. Available at www.ets.org/Media/Research/pdf/RM-97-01.pdf

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge, UK: Cambridge University Press.

Eggins, S., & Slade, D. (1997). *Analyzing casual conversation*. London: Cassell.

Freebody, P. (2013). Knowledge and school talk: Intellectual accommodations to literacy? *Linguistics and Education*, *24*, 4–7. doi:10.1016/j.linged.2012.11.004

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, *29*(3), 345–370. doi:10.1177/0265532211424479

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi:10.1016/j.jml.2007.11.007

Kintsch, W., & T. A. van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–94. doi:10.1037/0033–295x.85.5.363

Kroll, B. (1977). Combining ideas in written and spoken English: a look at subordination and coordination. In E. O. Keenan & T. L. Bennett (eds.), *Discourse Across Time and Space*. California, LA: University of Southern California, S.C.O.P.I.L. No. 5.

Lee, Y-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*, 131–166. doi:10.1191/0265532206lt325oa

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds), *Encyclopaedia of language and education* (*Vol. 7*: Language testing and assessment, pp. 121–130). Dortrecht, Netherlands: Kluwer. doi:10.1007/978–1–4020–4489–2_12

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185–207. doi:10.1111/lang.12117

Martin, J. R. (2013). Embedded Literacy: Knowledge as meaning. *Linguistics and Education*, *24*, 23–37.

Maton, K. (2013). Making semantic waves: A key to cumulative knowledge-building. *Linguistics and Education*, *24*, 8–22. doi:10.1016/j.linged.2012.11.005

Maton, K. (2014). *Knowledge and knowers. Towards a realist sociology of education*. Oxon and New York: Routledge.

Maton, K. (2016). Legitimation code theory: Building knowledge about knowledge-building. In K. Maton, S. Hood, & S. Shay (Eds.), *Knowledge-building: Educational studies in Legitimation Code Theory* (pp. 1–24). London: Routledge.

Matruglio, E., Maton, K., & Martin, J. (2013). Time travel: The role of temporality in enabling semantic waves in secondary school teaching. *Linguistics and Education*, *24*, 38–49. doi:10.1016/j.linged.2012.11.007

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. Mahway, NJ: Lawrence Erlbaum Associates.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*(1), 1–25. doi:10.1037/h0025327

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587. doi:10.1177/0265532209340192

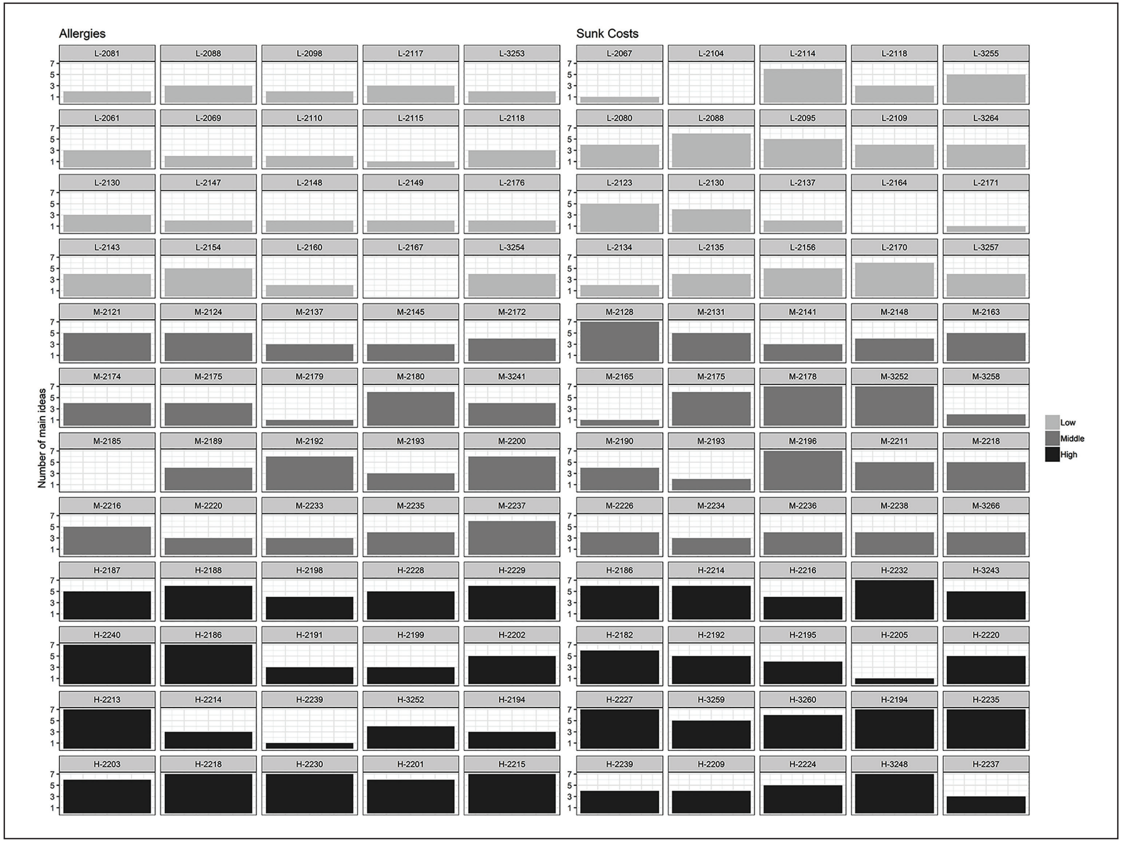
Plakans, L., & Gebril, A. (2012). A close investigation of source use in integrated second language writing tasks. *Assessing Writing*, *17*, 18–34. doi:10.1016/j.asw.2011.09.002

RStudio Team. (2016). *RStudio: Integrated Development for R. RStudio, Inc*. (Version 1.1.383). Boston, MA. Retrieved from www.rstudio.com/

RCore Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge, UK: Cambridge University Press.

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing*, *21*, 118–133. doi:10.1016/j.jslw.2012.03.004



Appendix. Individual reproductions of main ideas by topic and proficiency group.

The figure shows the number of source text main ideas reproduced by each participant in each proficiency group for each topic, Allergies and Sunk Costs. It highlights individual differences in each group across each task version.