Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks

Heng-Tsung Danny Huang
National Taiwan University, Taiwan

Shao-Ting Alan Hung
National Taiwan University of Science and Technology, Taiwan

Lia Plakans
The University of Iowa, USA

Abstract
Integrated speaking test tasks (integrated tasks) provide reading and/or listening input to serve as the basis for test-takers to formulate their oral responses. This study examined the influence of topical knowledge on integrated speaking test performance and compared independent speaking test performance and integrated speaking test performance in terms of how each was related to topical knowledge. The researchers derived four integrated tasks from TOEFL iBT preparation materials, developed four independent speaking test tasks (independent tasks), and validated four topical knowledge tests (TKTs) on a group of 421 EFL learners. For the main study, they invited another 352 students to respond to the TKTs and to perform two independent tasks and two integrated tasks. Half of the test takers took the independent tasks and integrated tasks on one topic combination while the other half took tasks on another topic combination. Data analysis, drawing on a series of path analyses, led to two major findings. First, topical knowledge significantly impacted integrated speaking test performance in both topic combinations. Second, the impact of topical knowledge on the two types of speaking test performances was topic dependent. Implications are proposed in light of these findings.

Keywords
Integrated speaking assessment, path analysis, topical knowledge, validity

Corresponding author:
Heng-Tsung Danny Huang, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 106, Taiwan.
Email: dannyhuang123@ntu.edu.tw
Introduction

Topical knowledge is a critical issue of concern with performance-based assessment. In the case of integrated language test tasks, a topic connects input and output as “the input that has been provided forms the basis of the response(s) to be generated by test-takers” (Lewkowicz, 1997, p. 121). In contrast to traditional independent language test tasks where test-takers produce answers without the benefit of any advance input (Brown, Iwashita, & McNamara, 2005), integrated language test tasks offer prior textual and/or aural input for test-takers to use to formulate oral or written responses. In the field of second/foreign language (L2) assessment, several well-established tests have been capitalizing on such tasks to assess test-takers’ writing and/or speaking skills. For instance, the TOEFL iBT assesses test-takers with tasks that integrate writing and speaking with reading and listening. The current research examined the impact of topical knowledge on integrated speaking task performance and explored how this impact would differ for independent speaking tasks.

Alexander, Schallert, and Hare (1991) define topical knowledge as “the intersection between one’s prior knowledge and the content of a specific passage or discourse” (p. 334). In the context of L2 assessment, this knowledge construct has been accorded considerable importance with theoretical frameworks that portray L2 speaking test performance and underlie L2 speaking test validation undertakings. However, topical knowledge has been shown to impose varying degrees of cognitive demands on L2 test-takers leading to performance variations. Jennings, Fox, Graves, and Shohamy (1999) have posited that topical knowledge, along with other constructs such as topic interest and perceived relevance, might induce topic effect, which influences test performance and causes unfair advantages or disadvantages to particular groups of test-takers.

Second language testing researchers have claimed that integrated test tasks may reduce topic effects on performance by providing prior input to equip the test-takers with the topical knowledge necessary to formulate responses. However, thus far, research examining topical knowledge in integrated speaking tests is scarce, leaving the claim that these tasks mitigate topic effects largely unsubstantiated. Additionally, while an extensive body of research has explored how topical knowledge influences L2 test performance, only a limited number of studies have centered on the assessment of L2 speaking skills. Research on the effects of topical knowledge on L2 speaking test performance has investigated the presence and degree of topical knowledge based on the test-takers’ academic backgrounds, learning experience, and gender.

Hence, to expand our understanding of topical knowledge in speaking assessments and, in particular, speaking assessment integrated with other skills, the present study measured topical knowledge with a set of custom-developed tests, examined its influence on integrated speaking test performance, and compared this influence with that on independent speaking test performance. The findings of this study inform L2 speaking test design by providing insight into integrated speaking tasks in terms of topical knowledge which can be useful for score interpretations and task development; they will also benefit topical-knowledge research endeavors by demonstrating the development procedure for an objective topical knowledge measure.
Integrated test tasks

The use of integrated tasks for language assessment has been bolstered by several perceived advantages. For example, integrated test tasks have been claimed to have authenticity because they are designed to simulate real-life language use tasks (Butler, Eignor, Jones, McNamara, & Suomi, 2000, p. 15), providing better predictive capacity, inducing more positive washback, and increasing learner motivation (Wesche, 1987). Additionally, integrated test tasks have been met with positive test-taker perceptions (Huang & Hung, 2010), which could affect performance and resulting scores (Stricker & Attali, 2010). A benefit most relevant to the present research is the topical knowledge made accessible to the test-takers through input materials. Test-takers bring varying degrees of topical knowledge to bear on the test tasks, and those weaker in relevant topical knowledge are presumed to be disadvantaged when they attempt their responses. Integrated language test tasks, by furnishing input materials ahead of time, may put the test-takers “on a more equal footing in terms of background knowledge” (Weigle, 2004, p. 30) and thus effectively mediate the potentially unfair impact of topic content on test performance (Read, 1990).

A review of the studies on integrated test tasks reveals that most have revolved around the assessment of writing skills, with much less attention to speaking skills. Among the studies on integrated speaking assessment, most investigated aspects other than the role of topical knowledge in the test-taking processes or test performance (e.g., Barkaoui, Brooks, Swain, & Lapkin, 2013; Crossley, Clevinger, & Kim, 2014; Lee, 2006; Swain, Huang, Barkaoui, Brooks, & Lapkin, 2009). Others, though foregrounding the importance of content, stopped short of examining whether topical knowledge has a part to play in how test-takers achieve content accuracy. For example, Brown et al. (2005) revealed that content constituted a salient conceptual category to which the raters directed attention in assessing performance on independent and integrated speaking test tasks on the new TOEFL project. Frost, Elder, and Wigglesworth (2011) analyzed the responses to the integrated listening-speaking task of an Oxford English language test and concluded that it is feasible to include a content dimension in the speaking assessment and “appropriate to consider accuracy of content as part of the construct of speaking ability” (p. 366). Jamieson and Poonpon (2013) expanded the holistic rating scale for the integrated speaking tasks on the TOEFL iBT into three analytic rating guides to offer more accessible rubrics for instructors and test-takers, one of which dealt specifically with the dimension of topic development. Taken together, these studies bring to light the pivotal importance of content accuracy in integrated speaking test performance; however, it remains a mystery whether and how content accuracy might hinge on the topical knowledge that the test-takers possess.

The only research effort currently available that touched on the role of topical knowledge in integrated speaking assessment pertained to Huang & Hung (2010) study that investigated the anxiety reactions and test-taker perceptions of an independent speaking task versus an integrated reading-speaking task. The results showed that the test-takers expressed an overwhelming preference for the integrated reading-speaking task because the input materials activated their relevant topical knowledge and also supplied the pertinent vocabulary to formulate their responses. Nonetheless, this study did not pursue the issue of how activated topical knowledge influenced integrated performance.
Topical knowledge

In the field of language testing, the role of topical knowledge has been emphasized in L2 theoretical frameworks proposed to describe language use/test performance (e.g., Bachman & Palmer, 2010) and oral test performance (e.g., Fulcher, 2003) as well as to guide validation efforts (e.g., Weir, 2005). For instance, in Bachman and Palmer’s (2010) model of language use, five test-taker characteristics (namely, topical knowledge, language knowledge, personal characteristics, strategic competence, and affective schemata) and the interactions among them constitute the essential components that underlie language use and test performance. In this model, topical knowledge reflects the knowledge repertoire with which individuals produce and interpret language.

In spite of such strong theoretical grounding, only a few L2 researchers have undertaken topical-knowledge studies in relation to the assessment of L2 speaking skills; their findings paint a complicated picture of the effects of topical knowledge on test performance. For instance, Douglas and Selinker (1993) recruited international teaching assistants (ITAs) to perform a general English proficiency test [the Speaking Proficiency English Assessment Kit (SPEAK)] and a field-specific version of SPEAK for mathematics majors (MATHSPEAK), and found that the mathematics majors performed better on MATHSPEAK than on SPEAK with the reverse being the case for the non-mathematics majors. Papajohn (1999) investigated performance of ITAs in chemistry who had performed both the SPEAK and the chemistry Taped Evaluation of Assistants’ Classroom Handling (TEACH); he discovered that topic changes led to variations in oral test performance. Lumley and O’Sullivan (2005) analyzed performances on the speaking component of an exit English proficiency exam and found that certain topics, such as horse-racing and basketball, led to bias that stood for or against a particular gender on oral test performance. Seedhouse and Egbert (2010) examined the topic development associated with the Speaking Test of IELTS and disclosed that some top-level test-takers experienced difficulty in responding to a specific topic regarding future changes in the relationship between education and work.

However, other studies have found topical knowledge to bear no relationship or a negative one with test performance. Smith (1989) prompted ITAs to undertake a general oral test (i.e., SPEAK) and one of the three field-specific tests depending on their majors. Her study uncovered that differences between the oral responses on the two tests were not statistically significant. Even more confusing, Douglas and Selinker (1992) administered three oral tests to Chinese-speaking ITAs in chemistry: SPEAK, CHEMSPEAK (a field-specific SPEAK for chemistry majors), and TEACH. Quantitative analyses indicated that the ITAs performed at a significantly lower level on CHEMSPEAK than on SPEAK, suggesting that familiarity with the subject matter somehow backfired and negatively impacted test performance.

The incongruence in the findings among these studies might have arisen from the procedures for defining topical knowledge. The studies reviewed identified the presence and degree of field-specific or topical knowledge by test-takers’ academic discipline (i.e., Douglas & Selinker, 1992, 1993; Papajohn, 1999; Smith, 1989), learning experience (i.e., Seedhouse & Egbert, 2010), or gender (i.e., Lumley & O’Sullivan, 2005). None of the researchers directly measured test-takers’ topical knowledge but made assumptions based
on the participants’ characteristics, which runs the risk of over- or under-estimation because there is no empirical evidence to support the assumption about knowledge (Dochy, Segers, & Duehl, 1999; Urquhart & Weir, 1998).

**The conundrum of topical knowledge in speaking assessment**

Based on this review of literature, two points can be gleaned about topical knowledge in speaking assessments. First, topical knowledge is recognized as a component underlying language use and spoken interaction (Bachman & Palmer, 2010; Fulcher, 2003; Weir, 2005). Second, research that has investigated topical knowledge in speaking test tasks has not yielded uniform results; some studies have supported the theoretical models and others have not. Despite this tentative relationship found between theory and research, language assessments, in general, do not seek to directly assess test-takers’ prior topical knowledge. One reason for this may be that it is not seen as ‘language’ per se, in a strictly linguistic sense. Language assessments seek to measure language ability or language use, not the knowledge a test-taker brings to the test on a topic, with the possible exception of language for specific purposes testing. If the test is not intended to assess topical knowledge but inadvertently does so, test fairness can be an issue. Jennings et al. (1999) contended that topical knowledge might offer unfair advantages/disadvantages to the test-takers. That is, when a test presupposes topical knowledge available only to some test-takers but not to others, the presence/absence of topical knowledge will introduce a bias for/against the test-takers as their performance reflects both the L2 ability being measured and the topical knowledge possessed. Perhaps for these reasons, testing researchers have hypothesized that topical knowledge might represent a potential source of construct-irrelevant variance (CIV) (e.g., Kunnan, 2000; Jennings et al., 1999). As defined by Messick (1995), CIV constitutes the “excess reliable variance … that affect[s] responses in a manner irrelevant to the interpreted construct” (p. 742) and would influence the judgments formulated for test performance (Messick, 1994). It could undermine “the accuracy of test score interpretations, the legitimacy of decisions made on the basis of test scores, and the validity evidence for tests” (Downing, 2002, p. 236).

This conundrum in language assessment – how to address topical knowledge but not measure it – can present challenges in defining the construct of speaking assessment and in providing meaningful score interpretations. One solution has been integrated assessment with the hope that, by providing test-takers with content through input material, topical knowledge will be leveled or controlled to lower its impact on scores. The current study explores the legitimacy of this assumption.

In this study, the speaking tasks used were designed for TOEFL iBT preparation; thus, the present researchers will draw on how topical knowledge is defined for that test purpose. In designing the framework for the speaking section of the TOEFL iBT, the target language use domain was narrowed to interactions in academic settings (Jamieson, Eignor, Grabe, & Kunnan, 2008). More specifically, the integrated tasks on the TOEFL “require the examinee to speak about a topic for which information had been supplied from another source” (p. 76). In a published report that informed the current TOEFL design, *TOEFL 2000: Speaking Framework A Working Paper* (Butler et al., 2000),
topical knowledge appears indirectly in a description of factors that impact success in speaking performance, as the third of three factors: “(c) the resources the individual brings to the interaction” (p. 2). The authors also recognize, in a footnote, that topic familiarity can be seen as a feature of task difficulty (p. 6). Therefore, the TOEFL iBT speaking tasks were designed with an awareness of and agreement with prior theoretical models positing that topical knowledge impacts speaking performance. However, this does not specifically position topical knowledge as part of the construct the test is intended to measure. The rubrics used to score the integrated and independent speaking tasks on the TOEFL (www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf) provide evidence that the intent is not to measure test-takers’ prior topical knowledge. Instead both rubrics include criteria for “topic development”: for the independent task, it is described in terms of sustained production on the topic and coherence, while, for the integrated tasks, it is tied into the accurate coherent presentation of content from the input materials. In both cases, prior topical knowledge is not mentioned directly. Returning to the conundrum of topical knowledge, for the TOEFL speaking tasks, while it is expected to impact performance, the test does not seek to directly measure topical knowledge in the test tasks or scores. This study uses tasks that follow the TOEFL format to consider whether topical knowledge is related to the scores on both integrated and independent speaking tasks to see how this factor might be understood in light of score interpretations and the difference between the two speaking task types.

The current study

This study was designed to contribute to the body of research on integrated skills assessment in connection to topical knowledge and provide evidence for the test-score interpretations based on integrated performance. Topical knowledge tests (TKTs) were developed to allow for examination of how topical knowledge impacted integrated speaking test performance. The study investigated whether independent and integrated speaking test performances differed in how they related to topical knowledge. The researchers operationalized the independent speaking test tasks as speaking-only tasks and the integrated speaking test tasks as reading–listening–speaking tasks and addressed the following research questions.

1. What is the relationship between topical knowledge and performance on integrated language tasks?
2. Does topical knowledge differentially impact performance on independent and integrated language tasks?

Method

The research process involved a preliminary stage and the main study. The preliminary stage designed and validated four TKTs. The main study drew on these four TKTs and eight speaking test tasks (i.e., four independent tasks and four integrated tasks) to collect and analyze the relevant data to address the research questions.
Participants

A total of 773 Taiwanese EFL learners participated in this research project, 421 in the preliminary stage and 352 in the main study. Among the participants in the preliminary stage, 405 finished at least one of the four TKTs. This group consisted of 127 males and 278 females and varied widely in age from 18 to 38. Furthermore, 323 were enrolled in a college at the time of study while the other 82 already possessed a college degree. Their fields of study varied broadly from English to Aquaculture. For the main study, 352 learners, 88 males and 264 females aged between 18 and 30, completed the speaking test tasks and TKTs. At the time of study, 19 were pursuing a master’s degree while the other 333 were working on a bachelor’s diploma. In terms of majors, they also came from a wide array of academic disciplines, but most (71%) specialized in foreign languages and majored in business-related fields of study in vocational high school.

Instruments

Speaking test tasks. Data for the main study were collected through eight semi-direct speaking test tasks: four integrated tasks and four independent tasks. The four integrated tasks, taking the form of reading–listening–speaking tasks, were chosen from commercially available TOEFL iBT preparation kits based on the criteria of being culturally, religiously, and controversially neutral. They were adapted slightly to end with a request for the test-takers to integrate information from the reading passage and the lecture into their oral production. These integrated tasks, as noted by the Educational Testing Service (ETS) specialists, assess the ability to “appropriately and intelligibly combine and convey key information from reading and listening texts representative of academic course content” (Pearlman, 2008, p. 246). The four integrated tasks, focusing on the topics of the euro, immunization, air transportation industry, and biofuels, invited test-takers to first read a passage, then listen to a lecture dealing with similar issues, and finally speak by integrating the passage and the lecture. Following the time allowance on the official TOEFL iBT, test-takers had 45 seconds for reading, roughly 90 seconds for listening, 30 seconds for preparation, and 60 seconds for speaking. The independent tasks were developed by the researchers with the goal of making it possible to counterbalance the topics in the official test administration (discussed below). These four tasks revolved around the same topics as the four integrated tasks and each required “test takers to draw upon their own ideas, opinions, and experiences” in response to a brief aural prompt (ETS, 2008, p. 16). As with such tasks on the official TOEFL iBT, test-takers spent 15 seconds in preparation and 45 seconds verbalizing their opinions.

In an attempt to reduce the potential bias of unfamiliarity with the speaking task formats, the researchers assigned the independent task on the euro and the integrated task on immunization as practice tasks to familiarize test-takers with the format. The tasks on air transportation industry and biofuels were assigned as the final tasks that provided the research data. Further, to eschew potential topic effects and to increase the generalizability of results, the researchers counterbalanced the topics for the final tasks (as shown in Table 1) by having half (\(n=173\)) take the final tasks in one topic combination (i.e.,
Combination A) and the other half \((n=179)\) complete another topic combination (i.e., Combination B).

Table 2 enumerates the characteristics of the passage and lecture of the two integrated speaking tasks for the two topic combinations. The two passages contained a similar number of words (i.e., 128 words and 123 words), a comparable amount of information (i.e., two information items each), and an approximately equal readability level (i.e., 50 points and 52 points). Likewise, the two lectures were similar in duration (i.e., 245 words and 249 words) and in the amount of information presented (i.e., three items and four items).

**Speaking rubrics.** The researchers rated the performance on the speaking tasks with reference to the independent and integrated speaking rubrics developed by the ETS (2008), which contain three key criteria: *delivery*, *language use*, and *topic development* (www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf). Specifically, *delivery* touches on pacing, intonation, pronunciation, and stress, and *language use* deals with grammar use, vocabulary choice, and idea cohesion. *Topic development* focuses on different aspects for the two types of tasks. For the independent tasks, it assesses the content relevance, idea elaboration, idea progression, and relationships between ideas. For the integrated tasks, *topic development* evaluates the completeness and specificity of the oral summary as required by the task. Based on these two sets of rubrics, each speaking task performance received a score ranging from zero to four for each criterion, garnering a total of three scores. Further, since this study administered the two speaking tasks to all participants (i.e., one independent task and one integrated task), each participant had six scores, three for each task.

**Topical knowledge tests.** In order to measure the participants’ approximate level of knowledge concerning the topics of the eight speaking test tasks (the euro, immunization, air transportation industry, and biofuels), the researchers constructed and validated four topical knowledge tests (TKTs) in the preliminary stage of this study. The construction phase of the TKT development consisted of two major steps. In the first step, the researchers searched relevant literature for a potential test format and selected the type of measure proposed by Valencia, Stallman, Commeyras, Pearson, and Hartman (1991), which predicts an individual’s topical knowledge from their familiarity with key terms on a topic. This prediction measure is designed to represent “an integration of three complementary areas of research related to comprehension – schema theory, knowledge of text structure and topical knowledge” and invites the test-takers to predict whether a list of ideas would appear in a passage on the specified topic (Valencia et al., 1991, p. 217).
This measure was chosen for the TKTs based on its relative ease of development, scoring, and piloting to capture topical knowledge more directly than prior research that used individual characteristics (e.g. major or gender) as indicators of topical knowledge.

In the second step, the researchers recruited six subject matter experts (SMEs) from the three academic fields (i.e., health, business, and biology) that relate to the four topics (i.e., the euro, immunization, air transportation industry, and biofuels) to generate the TKT items as well as to set the scoring key. Further, the validation phase of the TKT development comprised three primary steps. First, the researchers invited 421 Taiwanese college students or college graduates to take the four TKTs. Second, they performed a series of item analyses and exploratory factor analyses (EFAs) on the performance on these TKTs and retained or discarded items based on the results. Third, the researchers added distractors to and finalized the TKTs. These finalized TKTs for the euro, immunization, air transportation industry, and biofuels contained 15, 22, 16, and 13 items, respectively.

Each TKT required the participants to predict whether the given ideas would show up in a reading passage on one of the four topics. Their responses, when compared to those of the SMEs, would reveal the test-takers’ approximate level of knowledge of the topic. For instance, the TKT for biofuels called upon the participants to predict whether the 13 given ideas, including water, cancer, HIV, and so forth, would show up in the passage that dealt with the advantages and disadvantages of biofuels. To indicate their prediction for each idea, following Valencia et al. (1991), they could select yes, no, or maybe. The alpha coefficients of the TKTs on air transportation industry and biofuels, $\alpha = .74$ and $\alpha = .84$, respectively, have both reached an acceptable level of internal consistency ($air$ transportation industry: $n=351, k=13$; biofuels: $n=351, k=10$).

**Data collection**

To avoid clouding the picture of the relationship between topical knowledge and speaking test performance by unduly activating relevant topical knowledge, the main study collected research data in two stages set apart by one week.
participants gave informed consent for participation, filled out personal information sheets, and completed the four TKTs. In the second stage that occurred in computer labs, the participants first took the practice tasks and then performed the final speaking tasks. For the final speaking tasks, roughly half of the participants \((n = 173)\) completed Combination A, while the other half \((n = 179)\) finished Combination B (Table 1).

**Scoring**

The researchers scored the TKT answers and rated the speaking performances in the following manner. For the TKT answers, based on the practice in Valencia et al. (1991), they compared participants’ answers to those of the SMEs and assigned one point to every matched item, zero points for every mismatched item, and half a point to each *maybe*. Table 3 presents the scoring scheme for the TKTs.

In terms of the speaking performances, the first two researchers, both with extensive experience in EFL instruction and assessment, rated the performances on the speaking tasks with reference to the independent and integrated speaking rubrics introduced earlier. Prior to the official rating, they completed all four speaking tasks (i.e., the independent and integrated tasks on air transportation industry and the independent and integrated tasks on biofuels) to familiarize themselves with the prompts, examined the rubrics and level descriptors, and inspected the anchor performance for each level. Next, for each speaking task, they jointly rated five performances and reached agreed-upon scores for each performance. Subsequently, they independently rated another five performances and, once finished, discussed their rationale for awarding a particular score. Then, with reference to the two sets of speaking rubrics and these rater training experiences, they began rating the remaining performances on the four speaking tasks. Each performance was rated by both researchers. When the two scores awarded to the same aspect of a performance differed by more than two points (e.g., a four and a one), another colleague with a doctoral degree in language education would serve as the third rater and assess the performance and produce another score to be averaged with the closest one. The six Spearman’s rho \((\rho)\) estimates (i.e., three criteria for the independent performance and three criteria for the integrated performance), calculated based on the assigned ratings, varied slightly from .90 to .94, indicating an excellent level of rating congruence between the two raters.

**Data analysis**

Data analysis included preliminary analysis and primary analysis. The preliminary analysis scrutinized the data entry accuracy, searched and dropped the cases with missing

<table>
<thead>
<tr>
<th>Correct answer by the SMEs</th>
<th>Test-taker’s answer</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. The scoring scheme for the TKTs.
values, and tested major statistical assumptions. The primary analysis involved conducting path analysis, a statistical technique that “builds on ordinary multiple regression” (Retherford & Choe, 1993, p. 93) and aims to estimate “the magnitude and significance of hypothesized causal connections among sets of variables” (Stage, Carter, & Nora, 2004, p. 5). In addition to offering the benefits of multiple regression, path analysis further enables researchers to simultaneously evaluate all the hypothesized relationships among the modeled variables and allows for the assessment of direct and indirect effects of predictor variables on outcome variables as a way to cast light on the “the operative causal mechanism” (Olobatuyi, 2006, p. 12). Additionally, it encourages researchers to capitalize on path diagrams to establish and present an explicit theoretical causal model to explain the effects on outcome variables.

In this study, the researchers formulated the baseline path model to disentangle the relationship between topical knowledge and independent and integrated speaking test performance by following the guidelines by Keith (2006). This model was built from theories (e.g., Bachman & Palmer, 2010; Fulcher, 2003; Weir, 2005), relevant research (e.g., Douglas & Selinker, 1993; Huang & Hung, 2010; Lumley & O’Sullivan, 2005; Papajohn, 1999; Seedhouse & Egbert, 2010), time precedence, and logic. Figure 1 diagrammatically presents this baseline path model. In this model, independent – topical knowledge and integrated – topical knowledge represent the exogenous variables (i.e., independent variables) and independent performance and integrated performance reflect the endogenous variables (i.e., dependent variables). Also, the two endogenous variables each come with a disturbance term (i.e., e1 and e2) that denotes the unexplained variance and measurement error (Garson, 2014). Further, a single straight arrow represents the causal effect of an exogenous variable on an endogenous variable while a double-headed, curved arrow indicates the correlation between two variables (Lleras, 2005). As per the necessity to set the scale of measurement for the two disturbance terms, the researchers fixed the path linking each disturbance term to its endogenous variable to the value of one (Keith, 2006).

The researchers employed the AMOS 21.0 program with maximum likelihood estimation to assess this model by utilizing the correlation matrix of the four measured variables, namely, independent – topical knowledge, integrated – topical knowledge, independent performance, and integrated performance. Specifically, they evaluated the model–data fit based on the set of goodness-of-fit indices suggested by Kline (2005), namely, the χ² test statistic with its level of significance, the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). Because the χ² test has been found to have several limitations such as a high sensitivity to sample size (Hooper, Coughlan, & Mullen, 2008), this study further utilized the ratio of χ² statistic to degree of freedom (χ²/df) that provides a less restrictive measure (Kenny, 2015) as an additional model fit index. A fitting model should satisfy some or all of these model-fit criteria: a low and non-significant χ² (Hatcher, 1996), CFI > .95, RMSEA < .06, SRMR < .08 (Hu & Bentler, 1999), and a χ²/df ratio of less than five (Bollen, 1989; Marsh & Hocevar, 1985; Wheaton, Muthein, Alwirn, & Summer, 1977). Since the current study involved two topic combinations, the researchers assessed this baseline path model twice, once for each topic combination.
To investigate the second research question, the authors further developed a topicknowledge constraint model for each topic combination from its associated baseline path model, hereafter referred to as the constraint model (Figure 2). To establish each constraint model, they imposed an equality constraint on the two paths pointing toward the speaking performance constructs from the topical knowledge constructs to artificially equalize their strength (as symbolized by the letter $a$ assigned to the two paths in Figure 2). Since equality constraints would invariably deteriorate the model–data fit (Keith, 2006), if the equality constraint caused a non-significant deterioration of the model–data fit, it suggested that the two constrained paths then shared a comparable value and, by inference, topical knowledge did not differentially influence the independent performance and the integrated performance. Conversely, if the equality constraint significantly worsened the model–data fit, it indicated that the two constrained paths differed substantially from each other in magnitude and, by inference, topical knowledge did differentially impact the independent performance and the integrated performance. To assess the degree of model–data fit deterioration, the researchers compared the baseline model and the constraint model in terms of their relative model–data fit by way of the $\chi^2$ difference tests, the statistical technique that allows for the significance testing of the model–data fit deterioration introduced by equality constraints (Kline, 2005).

**Results**

**Preliminary analysis**

The researchers first ruled out the presence of out-of-range values (i.e., values that exceeded the maximum score) and missing values on the TKTs and ensured the clarity of
Huang et al.

recorded audio files. Next, they confirmed the absence of univariate outliers, detected and deleted one multivariate outlier, verified the distribution normality of the four modeled variables, and corroborated the non-existence of pairwise nonlinearity and multicollinearity. Taken together, the preliminary analysis eliminated only one case, retaining 351 cases for the primary analysis: 173 cases for Combination A and 178 cases for Combination B.

Path analysis

The baseline path models. The researchers performed a series of path analyses on the collected data with the two correlation matrices generated for the two topic combinations. Table 4 and Table 5 list these two correlation matrices, respectively, in juxtaposition with the mean values and standard deviations of the four modeled variables. As shown, for both topic combinations, all but one pair of variables shared a weak-to-strong positive correlation, indicating that, for each pair, higher values of one variable are associated with higher values of the other. For example, for Combination A, the positive, moderate correlation between integrated – topical knowledge and integrated performance implied that higher levels of topical knowledge about the integrated task were associated with higher levels of performance on the integrated task. Further, all of these positive correlations have reached statistical significance at the .05 level or greater.

Figure 3 and Figure 4 present the baseline path models for the two topic combinations annotated with the calculated parameter estimates and Table 6 summarizes the fit indices for these models. As shown by Figure 3 and Table 6, the path model for Combination A featured a CFI of .97 and an SRMR of .07, both of which met the pre-set criteria for a well-fitting model (i.e., CFI > .95; SRMR < .08). Further, although its $\chi^2$
**Table 4.** The correlation matrix for Combination A (n = 173).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1. Independent – topical knowledge</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Integrated – topical knowledge</td>
<td>.34**</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3. Independent performance</td>
<td>.37**</td>
<td>.07</td>
<td>1</td>
</tr>
<tr>
<td>4. Integrated performance</td>
<td>.29**</td>
<td>.34**</td>
<td>.63**</td>
</tr>
</tbody>
</table>

*Note:* Topics were biofuels (independent task) and air transportation industry (integrated task).

**Table 5.** The correlation matrix for Combination B (n = 178).

<table>
<thead>
<tr>
<th>Observed variables</th>
<th>Correlation</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1. Independent – topical knowledge</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Integrated – topical knowledge</td>
<td>.37**</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3. Independent performance</td>
<td>.38**</td>
<td>.14</td>
<td>1</td>
</tr>
<tr>
<td>4. Integrated performance</td>
<td>.16*</td>
<td>.22**</td>
<td>.73**</td>
</tr>
</tbody>
</table>

*Note:* Topics were air transportation industry (independent task) and biofuels (integrated task).

**Figure 3.** The estimated baseline path model for Combination A.
Topics were biofuels (independent task) and air transportation industry (integrated task).
statistic \[\chi^2 (2) = 7.25\] showed significance at the .05 level, its \(\chi^2/\text{df}\) ratio remained lower than the cutoff value of five \((7.25/2 = 3.63)\) and thus suggested an acceptable fit. Although its RMSEA (.12) exceeded the pre-set criterion (< .06), the 90% confidence interval of this RMSEA \([.04 – .23]\) included the criterion value; additionally, a possibility existed that this large RMSEA might have resulted from the small degrees of freedom associated with this path model \((df=2)\) since models with small degrees of freedom could possess artificially large RMSEAs even when correctly specified (Kenny, Kaniskan, & McCoach, 2015). Taken altogether, the model fit indices suggested the path model for Combination A provided a reasonable fit to the data.

A close analysis of the path coefficients, or the “numerical estimates of the causal relationships between variables in path analysis” (Olabatuyi, 2006, p. 41), reveals that in this path model, independent – topical knowledge made a large direct effect on independent
performance (β = .28) and integrated – topical knowledge also affected integrated performance substantially (β = .35) (based on the guidelines by Keith, 2006). In other words, these path coefficients suggested that topical knowledge exerted a substantial impact on speaking test performance on independent tasks and integrated tasks alike; namely, the stronger the level of topical knowledge, the better the speaking test performance, regardless of the task types. Further, the two topical knowledge constructs shared a moderate, positive correlation (r = .34) and the two performance constructs also correlated strongly and positively with each other, as demonstrated by the large correlation between their corresponding disturbance terms (r = .64). These correlation coefficients implied that topical knowledge about the independent task increased in proportion to the increase in topical knowledge about the integrated task; likewise, as independent performance increased, so did integrated performance. These path coefficients and correlation coefficients have all reached the level of statistical significance (p < .001).

As illustrated by Figure 4 and Table 6, the estimated baseline path model for Combination B emerged as an excellent representation of the data: χ² (2) = 1.47, p = .48; CFI = 1.00; RMSEA = .00; SRMR = .03; χ²/df = .74. A pattern similar to Combination A appeared in the significant relationships among the modeled variables: a large direct effect of independent – topical knowledge on independent performance (β = .33), a moderate direct impact of integrated – topical knowledge on integrated performance (β = .21), a moderate positive correlation between the two topical knowledge constructs (r = .37), and a strong positive correlation between the two performance constructs (r = .75).

The topical-knowledge constraint models. To respond to the second research question, the authors compared the effect of topical knowledge on independent and integrated performance statistically by examining the model–data fit between the baseline path model and the constraint model for both topic combinations. Prior to the comparisons, they evaluated the goodness-of-fit of the two constraint models and found that while the constraint model for Combination B could well reproduce the observed data, χ² (3) = 13.27, p < .05, CFI = .95, RMSEA = .14, SRMR = .07, χ²/df = 4.42, the constraint model for Combination A matched the collected data relatively unsatisfactorily, χ² (3) = 15.00, p < .05, CFI = .92, RMSEA = .15, SRMR = .08, χ²/df = 5. This unsatisfactory fit of the constraint model for Combination A revealed that the equality constraint would significantly harm the model–data fit for this model and, by inference, the two constrained paths in this model were not statistically comparable. As discussed immediately below, the χ² difference test performed to compare the baseline path model and the constraint model for this topic combination lent further support for this finding.

Table 7. Results of the χ² difference tests for the two topic combinations.

<table>
<thead>
<tr>
<th>Combination</th>
<th>Model</th>
<th>χ²</th>
<th>df</th>
<th>Δχ²</th>
<th>Δdf</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Baseline</td>
<td>7.25</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Constraint</td>
<td>15.00</td>
<td>3</td>
<td>7.75</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B Baseline</td>
<td>1.47</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constraint</td>
<td>13.27</td>
<td>3</td>
<td>11.80</td>
<td>1</td>
<td></td>
<td>&lt; .01</td>
</tr>
</tbody>
</table>
Table 7 delineates the results of the $\chi^2$ difference tests computed to compare the baseline path model and the constraint model in terms of their fitness to the data for both topic combinations. As shown, the results for Combination A, $\Delta \chi^2 (1) = 7.75, p < .01$, suggested that the constraint model did exhibit a significant loss of model-fit as compared to the baseline path model; an identical picture emerged for the results for Combination B, $\Delta \chi^2 (1) = 11.80, p < .01$. These results implied that, for both topic combinations, the values of the paths from the two topical knowledge constructs to the two performance constructs differed from each other in a significant way. Topical knowledge indeed differentially affected the independent performance and the integrated performance. Considered in conjunction with the baseline path models, these results demonstrated that, for Combination A, topical knowledge exerted a stronger impact on integrated performance than on independent performance. However, this pattern of impact reversed for Combination B; that is, topical knowledge exercised a larger influence on independent performance than on integrated performance.

**Discussion**

This study modeled the relationship between topical knowledge and integrated performance and explored whether topical knowledge affects independent and integrated performance differently. Path analyses revealed that (1) topical knowledge significantly affected both independent and integrated performance across the two topic combinations and (2) topical knowledge indeed influenced independent and integrated performance differently.

*The impact of topical knowledge on L2 speaking test performance*

The findings that topical knowledge constituted a statistically significant determinant of L2 speaking test performance, be it independent or integrated performance, confirmed that this factor is an important component of L2 speaking performance and effectively substantiated the theoretical frameworks reviewed previously that implicate or emphasize the essential role of topical knowledge in L2 performance (e.g., Bachman & Palmer, 2010), oral test performance (e.g., Fulcher, 2003), and test validation (e.g., Weir, 2005). On the other hand, the finding may also lend empirical support to the claim of topical knowledge acting as a source of CIV (e.g., Kunnan, 2000; Jennings et al., 1999) as it might offer an unfair (dis)advantage to the test-takers, when it is not part of a test’s construct or the purpose for which the test was designed.

Furthermore, the findings demonstrate that integrated assessment tasks are not immune to the influence of prior topical knowledge on scores. Namely, the claim in the literature on integrated assessment (Read, 1990; Weigle, 2004) suggesting that the input materials decrease the effect of topical knowledge by minimizing the unfair (dis)advantage was not confirmed. In fact, the current findings suggest otherwise; pre-task topical knowledge still plays a significant role in integrated performance. The two baseline path models (Figure 3 and Figure 4), by identifying a moderate or large effect of topical knowledge on integrated performance, provided counter-evidence for this claimed benefit and confirmed the presence of an advantage accorded to those stronger in pre-task topical knowledge.
Collectively, these findings combined to provide evidence showing that topical knowledge is a significant source of CIV for integrated assessment, whose interpretation is not intended to include content knowledge. It is interesting to note that this finding is also in contrast to reported test-takers’ perceptions in a study where participants responded positively to integrated reading-speaking assessment because they believed its input materials activated topical knowledge to generate answers, neutralizing the impact of any lack of pre-task topical knowledge on their performances (Huang & Hung, 2010).

The impact of topic knowledge on integrated task performance may be described as a cumulative advantage for test-takers possessing more topical knowledge to begin with. DiPrete and Eirich (2006) define cumulative advantage as “a general mechanism for inequality across any temporal process … in which a favorable relative position becomes a resource that produces further relative gains” (p. 271). Integrated tasks require reading and listening to generate an oral response, and topical knowledge benefits reading and listening processes (e.g., Carrell, 1987; Chiang & Dunkel, 1992). These benefits also exist for speaking performances. Thus, the cumulative advantages from pre-task topical knowledge starts with an increased comprehension of the reading and listening input, which extends further to the speaking performance. Therefore, while the input offered by integrated tasks provides content for test-takers’ oral responses, it might be more accessible to test-takers who comprehend input better due to pre-task topical knowledge. In other words, by supplying input, the integrated tasks might have somehow promoted the rich-get-richer phenomenon and allowed the presence of relevant topical knowledge to take on an even more facilitative/debilitative role in performance variations.

The differential impact of topical knowledge on independent and integrated performance

While topical knowledge indeed differentially impacted independent and integrated performance, this impact was different depending on the topic. Topical knowledge affected integrated performance more strongly in Combination A but had the opposite effect in Combination B where it had a greater impact on independent performance. This result presents a somewhat confusing circular hypothesis – there is a topic effect on the effect of topical knowledge. A closer review of Table 1 reveals that the two performances that sustained a stronger effect of topical knowledge (i.e., the integrated performance in Combination A and the independent performance in Combination B) dealt with the topics of air transportation industry, which relate to the field of business. Considered alongside the fact that most of the participating students (71%) majored in business-related fields when they studied in vocational high school, this finding suggests that when test-takers have accumulated more topical knowledge prior to taking a speaking test task, their performance will benefit from this knowledge, regardless of the task format. Conversely, when the test-takers knew relatively less about the topic in advance of taking the speaking test task, their performance suffered, no matter which form the task took.

Potentially, this finding adds more evidence to contradict the claim of integrated skills tasks stating that input materials reduce the effect of topical knowledge on performance. It provides counter-evidence by showing that although integrated language test tasks have provided topical knowledge for test-takers to produce their responses (Weigle,
2004), those stronger in the initial level of topical knowledge still perform better than their counterparts with less relevant knowledge to begin with. This finding seems to allude to the possibility that offering input materials alone might not mediate the impact of topical knowledge on test performance (Read, 1990), and that further test design modifications appear necessary to minimize this potentially unfair impact. This speculation needs further research study.

Limitations and implications

The findings of this study should be interpreted in light of several limitations inherent in its research design. First, the test takers in this study performed the speaking test tasks voluntarily rather than as an assessment of consequence, which may affect the effort they put into their performances. Although the current researchers attempted to increase participant motivation through repetitively emphasizing the utility of this study and providing feedback on their performance for improvement, it is not known exactly what the effects of these efforts were. Second, the researchers gathered the speaking performance data using test tasks derived from commercially available TOEFL iBT preparation kits. While these kits claim to include tests constructed to simulate the official TOEFL iBT, they offered little description of the development and validation processes. In order to monitor and control limitations as much as possible, the design of this research included such steps as conducting rigorous piloting of the tasks and checking for outlier performances.

The findings of this research hold important implications for L2 assessment research, as test developers must be vigilant in exploring ways to minimize topic effects for tests that do not include topical knowledge in the constructs they intend to assess. While this is not a new challenge, this study’s findings suggest that the solution may not lie simply in providing input materials and adopting integrated task formats, and that there is a need to continue investigating viable solutions for all tasks, including those that provide input materials. For example, studies could examine topic effects in relation to characteristics of source texts (i.e., length, number, difficulty, or modality). Researchers might consider how including content accuracy on the rating scales may augment, if not exacerbate, the impact of topical knowledge on integrated performance. In sum, the present study provides counter-evidence to the assumption that integrated tasks will overcome the problem of construct-irrelevant topic effects and asserts the need of more research in this area.

Conclusion

As detailed by Cumming (2014), five major challenges confront integrated skills assessment: presence of task dependencies, lack of diagnostic information on low-scorers, absence of a well-defined genre, requirement for a threshold level of proficiency, and inappropriate source borrowing. The findings of this study have added to these challenges the potential topic effects compounded by these tasks, which are a concern for test-score interpretations. The results of this research provide insight into the conundrum of topical knowledge with the finding that it has an influence on language use and speaking performances on both integrated and independent speaking test tasks. This conclusion does not resolve the issue of how or if it should be manifest in constructs for speaking
assessment, but emphasizes that the role of topical knowledge deserves critical and con-
stant attention in defining constructs, creating tasks, and designing scoring schemes.
While integrated assessment tasks hold promise and address some limitations of inde-
dependent tasks, they are still subject to the impact of what test takers bring to the task.
Therefore, L2 scholars and practitioners, while taking into account the promises inte-
grated skills assessment has to offer, might do well to bear closely in mind these chal-
lenges as they attempt to evaluate L2 performance with this assessment procedure.

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 study was undertaken based on part of the research data collected
for the first author’s doctoral dissertation sponsored by the TOEFL Small Grants for Doctoral
Research in Second or Foreign Language Assessment program at ETS. Thus, the authors would
like to express gratitude to this program for its financial support.

Notes
1. This study drew in part on the research data collected and used in the first author’s doctoral
dissertation; yet, only the components directly relevant to the current study are described
here.
2. Since the researchers did not need to counterbalance the topics for the practice tasks, they left
out the independent task on immunization and the integrated task on the euro from the official
test administration. In other words, the main study collected data only on six speaking test
tasks, namely, three independent tasks and three integrated tasks.
3. The items remaining in the final TKTs from the EFAs all happened to be items with no as the
correct answer, so the researchers added three yes items as the distractors to each TKT in
order to hide this answer pattern.
4. The researchers did not analyze the responses on the TKTs for the euro and immunization
since their corresponding tasks only served the warm-up function and did not produce the
formal research data.

References
in learning and literacy talk about knowledge. Review of Educational Research, 61,
315–343.
Press.
Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers’ strategic behaviors in inde-
taker performance on English-for-academic-purposes speaking tasks (TOEFL Monograph


