1 Variability in speaking assessment and the role of topic

Introduction

Topics are often used as a key speech elicitation method in performance-based assessments of spoken language. They thus constitute an important area for validity enquiry. For instance, are different topics ‘equivalent’ or ‘parallel’? Can some topics bias against or favour individuals or groups of individuals? Does background knowledge of topic have an impact on performance? Might the content of test taker speech affect their scores – and perhaps more importantly, should it?

In performance-based assessments of speaking, a common practice for eliciting speech is to engage test takers with a topic or range of topics. To address these topics, test takers often draw on their topic-related background knowledge (BK), which generally serves as an information base for performance to be built upon (Bachman and Palmer 1996). To illustrate, a test taker might be asked to talk about an important festival, a newspaper article they have read, or a recent holiday. Test takers would then need to draw on their knowledge and experiences of the topics as well as their language skills in order to formulate a response.

In administering different topics to test takers, there is an underlying assumption that (all other things being equal) the speaking tasks are of equivalent levels of difficulty regardless of the choice of topic and can thus be considered ‘parallel’. What logically follows is a second assumption that the individuals’ degree of topic-related BK does not have a significant influence on their test results. Evidence to the contrary, however, may suggest that a validity threat has been introduced to the test owing to the influence of the construct-irrelevant factor of BK. Moreover, test fairness may also be compromised if it is shown that individuals or groups of individuals have been favoured or biased against as a function of their BK (Jennings, Fox, Graves and Shohamy 1999). Given that the results of tests, particularly large-scale standardised ones, are used to make decisions about test takers, these validity concerns become critical. A review of the literature on the effects of topic and BK of topic on performance, however, points to a need for more empirical research on these issues, particularly in the context of speaking.
On Topic Validity in Speaking Tests

This volume reports on an empirical research study investigating the role of topic and BK of topic in the Speaking module of IELTS (International English Language Testing System); an established and widely used, face-to-face second language (L2) speaking test. It draws on original data as well as insights from empirical and theoretical research to address some of the questions and issues raised so far. By grounding the research in the real-world assessment context of IELTS, this volume allows for an exploration of topic validity against the backdrop of one of the world’s most high-stakes English language tests.

Variability and spoken performance

Variability in L2 spoken performance assessment is an area of significant interest and debate from both a theoretical and an empirical standpoint. A number of factors other than the speaking ability in question may have the potential to influence performance in an assessment context (McNamara 1996, Milanovic and Saville (Eds) 1996).

McNamara’s (1996:86) model of proficiency, for example, illustrates the complexities involved in performance assessment, while Kunnan (1995:6) lists several test taker characteristics such as age and gender as well as other social, cognitive, and psychological factors such as cultural background, aptitude, and learning styles as having a potentially ‘critical influence’ on L2 performance (see also O’Sullivan 2000 for a synthesis of the literature on test taker characteristics). Other factors that have been identified in the literature as potentially exerting an impact on performance include (but are not limited to): characteristics of the tasks and processing conditions (De Jong and Vercellotti 2016, Luoma 2004, Skehan and Foster 1997); characteristics of the interlocutor(s) and/or rater(s) – such as gender, proficiency level, and personality (Nakatsuhara 2011, O’Sullivan 2000); raters’ degree of harshness or leniency when rating (Eckes 2009, McNamara 1996, McNamara, Knoch and Fan 2019, Yan 2014); raters’ approaches to scoring and interpretations of rating scale criteria (Baker 2012, Cumming, Kantor and Powers 2002, Lumley 2002, Milanovic, Saville and Shuhong 1996); characteristics of the rating scales (Barkaoui 2007, 2010); degree of acquaintanceship between interlocutors (O’Sullivan 2002); and lastly, socially constructed phenomena such as power relations (Shohamy 2001).

The reason why variability in performance assessment is so critical is that the resulting score from a test is used for making inferences about test takers’ abilities and for making important decisions about them. A score, however, can be ‘attractively simple’ and yet ‘deceptive’ (McNamara 2000:55). To illustrate, a score of 5 awarded by a harsh rater on a difficult task is meaningfully different from a score of 5 awarded by a lenient rater on an easy task. Extending the argument, we can say that for a cultural topic such as the
Variability in speaking assessment and the role of topic

Mexican Día de Muertos (Day of the Dead), a score of 5 may have different meanings for a test taker who is familiar with this event from one who has little BK or experience of it.

From a test validity perspective, it is important to monitor the potential effects of factors extraneous to the ability being measured, that is, construct-irrelevant variables (McNamara 2000), and to consider ‘other plausible rival interpretations of score meaning’ (Messick 1996:246). This now brings us to the parameters of interest in the empirical research reported in this volume: topic – as a task characteristic – and BK of topic – as a test taker characteristic.

Speaking task characteristics: Focus on topic

Spoken performance, as discussed in the section above, can be influenced by several parameters such as characteristics of the task, the test taker, the interlocutor, the rater, the rating scale and criteria, as well as the interactions between them (McNamara 1996). Speaking test tasks play a pivotal role in assessment; they serve as a link between test takers’ underlying abilities and subsequent performance through eliciting samples of speech (Fulcher 2003). Speaking tasks can be defined as ‘activities that involve speakers in using language for the purpose of achieving a particular goal or objective’ within particular settings (Bachman and Palmer 1996:44). Moreover, by manipulating task characteristics and administration conditions, test designers can direct and influence candidates’ performance to a certain extent (Luoma 2004). Of relevance here is a distinction made by Brown, Anderson, Shillecock and Yule (1985) between ‘chatting’ and ‘information-related talk’ as representing two ends of a continuum in respect of the purposes of ‘talk’. Chatting is viewed as a predominantly social activity that involves ‘finding a fluid stream of topics that the speakers find sufficiently interesting to take up, and on which they can find a shared angle’ (Luoma 2004:22). These topics are not necessarily discussed in great depth. At the other end of the continuum is ‘information-related talk’ described as ‘speech aimed at transferring information on a particular topic’ and is the one more often used in assessment contexts (Luoma 2004:23). The information-oriented nature of speaking tasks in assessment contexts thus highlights the importance of the task topic and the test takers’ information about the specific topic.

Topic features prominently in models of language use and task-based performance. In Bachman and Palmer’s (1996) influential model of communicative language ability, topic is identified as a component of the language of test task ‘input’ with input described as what the test takers are supposed to process and subsequently respond to. The topic component carries information in the input of the task and can be ‘personal, cultural, academic, or technical’ (Bachman and Palmer 1996:53). The potentially facilitating or impeding role of topical knowledge in relation to task topics...
On Topic Validity in Speaking Tests

is commented on by the authors, who argue that ‘certain test tasks that presuppose cultural or topical knowledge on the part of test takers may be easier for those who have that knowledge and more difficult for those who do not’ (Bachman and Palmer 1996:65). Weir (2005:76) also voices a concern that different topics may elicit ‘responses that are measurably different’. Illustrative examples from the literature include academic and technical topics such as the ‘natural virus’ topic for medicine majors and a ‘computer virus’ topic for computer science majors (Bei 2010), and cultural topics such as the Moon Festival being considered more familiar for Chinese learners compared to St Patrick’s Day (Li et al 2017) (see Chapter 2 for more examples and details of these studies).

Whereas topic is viewed as an important characteristic of tasks in general, it is reasonable to assume that its salience may also be affected by task type. Two task types are of particular relevance here: integrated speaking tasks and independent or stand-alone speaking tasks. Integrated speaking tasks are defined as tasks that ‘involve combinations of reading, listening and/or writing activities with speaking’ (Luoma 2004:43) and require test takers to speak about a topic for which information has been provided from other sources (Jamieson, Eignor, Grabe and Kunnan 2008). The Internet-based Test of English as a Foreign Language (TOEFL iBT) integrated speaking task, for example, requires the candidate to first read a passage about a campus-related topic, to then listen to a conversation about the same topic, and to subsequently prepare a response that summarises and brings together the information from the two input sources.

In contrast, independent tasks require test takers to ‘draw on their own ideas or knowledge [in order] to respond to a question or prompt’ (Brown, Iwashita and McNamara 2005:1). An example of a TOEFL iBT independent speaking question is:

Some people think it is more fun to spend time with friends in restaurants or cafes. Others think it is more fun to spend time with friends at home. Which do you think is better? Explain why.

Unlike the previous integrated speaking task example, the test taker is not supplied with any additional reading or listening input to engage with in addressing this prompt.

These two task types have been compared on several aspects such as degree of authenticity (particularly in academic contexts), content coverage, generalisability, cognitive processing demands on test takers, and reliability of ratings (Barkaoui, Brooks, Swain and Lapkin 2012, Lee 2006, Luoma 2004).

1 www.ets.org/toefl/test-takers/ibt/about/content/speaking/q2-integrated-transcript
2 www.ets.org/toefl/test-takers/ibt/about/content/speaking/q1-independent-transcript
Of particular relevance to this discussion are the different approaches to addressing topic-related BK in these task types. In integrated tasks, an attempt is made to minimise and/or mediate the (negative) impact of BK through the provision of input in the form of reading and listening texts. Independent tasks, on the other hand, require test takers to draw on their own BK. This absence of input in independent tasks has been criticised for not allowing an ‘equal footing’ (Weigle 2004:30) for test takers who bring varying degrees of BK to a test and for the restriction of topics to ‘fairly bland’ ones (Brown et al 2005:1). In this light, integrated tasks are viewed as ‘promoting equity or fairness’ (Huang 2010:4). We can also argue that by providing test takers with the necessary ideas for responding to a topic (instead of asking them to generate ideas and rely on their own BK), the cognitive demand of tasks can be reduced to a certain extent (Field 2011). As Jennings et al (1999) caution, this is not to say that a topic effect does not exist in integrated tasks but that the impact is likely to decrease owing to the provision of input. By the same token, it is plausible to assume that any effects of topic and BK of topic are likely to be manifested more markedly in independent tasks. Recent research, however, suggests that integrated tasks may not be necessarily ‘immune to the influence of prior topical knowledge on scores’ and that BK can be a ‘significant determinant’ of speaking test performance regardless of task type (Huang, Hung and Plakans 2018:43).

**Test taker characteristics: Focus on background knowledge of topics**

A discussion of task topics is inextricably linked to test takers’ BK of topics. BK is referred to in the literature under different terms such as content knowledge, prior knowledge, schematic knowledge, topical knowledge, and world knowledge. These terms are often used interchangeably although there has been a recent move towards establishing the nuances between the different terms (see for example Banerjee 2019 and O’Reilly and Sabatini 2013). Broadly speaking, a facilitative role for BK on performance has been posited in the theoretical literature.

A central role, for example, has been ascribed to BK in language comprehension as formalised in schema theory (Bartlett 1932, Carrell and Eisterhold 1983, Rumelhart 1980).

Text, any text, either spoken or written, does not by itself carry meaning. Rather according to schema theory, a text only provides directions for listeners or readers as to how they should retrieve or construct meaning from their own previously acquired knowledge. This previously acquired knowledge is called the reader’s *background*...
knowledge, and the previously acquired knowledge structures are called schemata (Carrell and Eisterhold 1983:556; emphases in original).

A facilitative role for topic familiarity on performance has also been suggested in Skehan’s (1998) framework of task processing conditions. It is hypothesised that the more familiar a topic is to an individual, the less cognitive load it poses by providing ‘easy access to information [which] should make only limited demands on attention, allowing material to be assembled for speech more easily and with greater attention to detail’ (Skehan 2001:175). In a similar vein, Robinson’s triadic componential framework (2001) views prior knowledge as a cognitive complexity dimension where unfamiliar tasks and those for which individuals’ prior knowledge is lower can increase task complexity, leading to ‘a depletion of attentional and memory resources’ (Robinson 2001:308; emphasis in original) affecting the accuracy and complexity of performance.

In the field of language assessment, Bachman and Palmer (1996:65) define topical knowledge as ‘knowledge structures in long-term memory’ that can have a substantial effect on performance. Topical knowledge features as one of the five main components of Bachman and Palmer’s (1996) model, inseparably linked to all instances of language use, as it ‘provides the information base that enables them [individuals] to use language with reference to the world in which they live, and hence is involved in all language use’ (Bachman and Palmer 1996:65). BK is also often considered in relation to potential sources of test bias where the test task ‘contains content or language that is differentially familiar to subgroups of test takers’ (O’Sullivan and Green 2011:61).

Despite the pronounced role attributed to BK of topics on performance from a theoretical standpoint, the results of empirical studies on the subject are often mixed and inconclusive (see Chapter 2). One possible reason is the various ways in which BK has been operationalised in the literature, for example, as knowledge related to academic field of study (Clapham 1996), cultural background (He and Shi 2012), gender (Lumley and O’Sullivan 2005), religious background (Markham and Latham 1987), and personal interest in topics (Jennings et al 1999). Furthermore, the majority of empirical studies have explored the effects of BK on reading and listening comprehension with fewer studies focusing on the performance skills.

Only a handful of studies have exclusively examined topic and BK effects on speaking. This is surprising, as the case for speaking is arguably stronger than the other skills; the online nature of speaking requires almost instantaneous access to BK for the spontaneous generation of ideas necessary for addressing independent topic-based tasks. Given the importance attributed to the two associated factors of topic and BK of topic, a closer examination of these two variables on speaking performance is warranted.
Variability in speaking assessment and the role of topic

This is particularly so in assessment settings where topic plays a critical role in eliciting speech, as is the case in the IELTS Speaking test. In the next sections, I will provide more details of the research context and illustrate the centrality of topics in the test as the main ‘driver’ of speech.

Research context

What is IELTS?

Research in language assessment, more often than not, is linked to specific exams, testing instruments, and validation efforts. By grounding research in real-world assessment contexts, results of studies can influence and shape testing practices with the potential to impact a large number of individuals and organisations, particularly in the case of large-scale standardised tests.

IELTS is one of the world’s most popular English language tests used for study, migration, or work. It has a candidature of over 3.5 million per year and is taken in 1,600 test centres in more than 140 countries around the world.

IELTS has two modules: Academic and General Training. The Academic module is designed to assess English language proficiency for those applying for higher education or professional registration. General Training is used to measure proficiency for more practical use in social contexts and is used for migration as well as other purposes such as training, secondary education, and work in English-speaking environments. Both modules have four papers covering the skills of listening, reading, writing, and speaking. The Listening and Speaking papers are common across the modules with the subject matter of the Reading and Writing sections as the main differentiating factor between the two.

IELTS scores are reported on a nine-band scale from non-user (a score of 1) to expert user (a score of 9). There is no pass or fail in IELTS on the grounds that ‘the level of English needed for a non-native speaker student to perform effectively varies by situation and institution’. Some guidance on overall IELTS scores based on the linguistic demands of academic curricula is provided though this is not designed to be prescriptive and instead, organisations and institutions are encouraged to set their own minimum scores on the basis of their specific requirements.

3 www.ielts.org
4 www.ielts.org/about-ielts/ielts-test-types
5 www.ielts.org/-/media/pdfs/ielts-test-score-guidance.ashx
6 For more information on the test as well as the latest research on IELTS, visit the IELTS website (www.ielts.org). For information about the historical development of IELTS, the interested reader is referred to Davies (2008). The volume chronicles the evolution of IELTS against the historical backdrop of academic English language proficiency testing within
IELTS can be safely categorised as a high-stakes test: ‘high-stakes decisions are major, life-affecting ones where decision errors are difficult to correct. Because of the importance of their effects, the costs associated with making the wrong decision are very high’ (Bachman 2004:12). Putting test takers at the heart of assessment, Shohamy (2001:102) defines the criterion for a high-stakes test as ‘whether the results of the test lead to detrimental effects for the test takers’. Cronbach (1988:6) argues that ‘tests that impinge on the rights and life chances of individuals are inherently disputable’, and they should thus be accessible to critical reflection and dialogue within the language testing (LT) community (Fulcher and Davidson 2007). It is this very high-stakes nature of IELTS that demands rigorous research on every aspect of the test in order to ensure its validity and fairness.

The IELTS Speaking test

The IELTS Speaking test (IST) is a face-to-face oral interview between a test taker and a certified IELTS examiner. The interview lasts between 11 and 14 minutes and is recorded. There are three main parts in the IST. Following an introduction, Part 1 is an Interview task (also known as Information Exchange task) where the examiner poses a series of questions on some general and familiar topics. In Part 2, the Individual Long Turn, the candidate is presented with a printed task card, which requires an extended talk on a specific topic for 1 to 2 minutes. The task card includes points that the candidates can cover in their monologue. Prior to the monologue, the candidate is given 1 minute of silent planning time as well as pencil and paper for making notes. The third part of the test, which lasts about 4 to 5 minutes, is termed a Two-way Discussion where the examiner poses several questions on more abstract topics, which are thematically linked to the topic in Part 2. This final part aims to provide the candidate with an opportunity to discuss these more abstract themes and topics.

Part 1 consists of three multi-question topic sets (or topic frames). The first frame involves a general topic such as work or studies and the remaining two sets involve other familiar topics. Part 2 consists of one topic-based monologue task. Similar to the Interview task, the Two-way Discussion often consists of two topic sets. The topic sets for Parts 1 and 2 are randomly

the UK higher education system. In his reflections, Davies (2008) draws attention to the ongoing conflict between theoretical stances in defining constructs of language proficiency within applied linguistics on the one hand, and the practical demands of standardised large-scale assessment on the other. This is a point worth emphasising, as practicality concerns – particularly those associated with large-scale assessment – are often neglected in research studies.

7 For a comprehensive historical overview of the IST see Nakatsuhara (2018).
Variability in speaking assessment and the role of topic assigned to test takers and the topic sets from Part 3 of the test are thematically linked to the Part 2 topic.

The IST is scripted and standardised, and examiners are given detailed instructions to follow in terms of test administration and management. The reliability of test delivery is achieved through the use of an Examiner Frame or Script which is ‘a script that must be followed’ (IELTS Examiner Training Material cited in Seedhouse and Harris 2011:72; emphasis in original) and which is designed to carefully delineate the examiner’s role in the interaction with the candidate, ‘[guiding] test management’ as it progresses through different parts of the test (Taylor 2007b:187).

The construct underlying IST is communicative (spoken) language ability. The test is designed to assess a wide range of skills which correspond to the different parts of the test and aims to elicit ‘the ability to communicate opinions and information on everyday topics and common experiences and situations by answering a range of questions; the ability to speak at length on a given topic using appropriate language and organising ideas coherently; and the ability to express and justify opinions and to analyse, discuss and speculate about issues’.

The IELTS speaking scale is a nine-band analytic scale consisting of four criteria: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation. IELTS examiners evaluate candidate performances throughout the test and award a final score once the candidates have left the session. The IELTS examiner thus serves the dual role of interlocutor and rater. The IST is single-marked and test reliability is assured through examiner training and certification, standardised procedures, a sample monitoring system, a ‘jagged profile’ system, and routine validation of task, candidate and examiner performance (Taylor 2007a:29).

**Topic as main ‘driver’ of speech in the IST**

Let us now consider the IST and its structure with a view to contextualising previous discussions on topic effects on performance with reference to specific test features.

In all three parts of the IST, speech is elicited by means of different topics, from more familiar ones in Part 1 to more abstract themes in Part 3. This gradation is designed to cater to the IELTS candidature who can vary widely in their speaking ability levels. The familiar/unfamiliar and concrete/abstract continua are used to span this divide in terms of demands on candidates and scope for sufficient production to allow for meaningful evaluation.

---

On Topic Validity in Speaking Tests

Based on their findings from a conversation analysis of transcribed IELTS spoken performances, Seedhouse and Harris (2011) identify topic as an integral component of the IST and as a vehicle for organising talk, driven almost exclusively by the examiner script. Referring to the organisation of talk in the test as a ‘topic-based Q-A adjacency pair’, the authors illustrate, in their examination of representative performances, how, in contrast to normal conversation, ‘topic is always introduced by means of a question’ (Seedhouse and Harris 2011:69). The two elements of the adjacency pair involve a question posed by the examiner to which the candidate has to respond and a ‘topic’ element which calls for the development of a specific topic (Seedhouse and Harris 2011:83). The findings also suggested the primacy of the Q-A element over the topic element in those instances where the two do not co-occur as ‘candidates can answer questions without developing topics’ (Seedhouse and Harris 2011:83). Put differently, the provision of a response, which might only minimally answer a question, may be perceived by candidates to be more important than elaborating on a response by means of topic development. In these cases, the information-transfer function of the test takes precedence over its main purpose, that is, speech generation for the purposes of evaluation. Drawing on this research amongst others, Seedhouse (2018) views topic as a fundamental construct within the IST exhibiting what he calls ‘a dual personality’ (2018:114): ‘topic-as-script’ and ‘topic-as-action’. The former refers to ‘the homogenised topic which examiners give to candidates’ and the latter refers to the ‘diverse ways in which candidates talk a topic into being’ (2018:114) which could subsequently impact performance scores.

Topic, therefore, constitutes the main vehicle for driving talk in the IST and yet, a specific topic development or content-oriented criterion is surprisingly absent in the IST Band Descriptors, which might explain why test takers may not always elaborate on topics (Seedhouse and Harris 2011). Topic is referred to under the Fluency and Coherence scale to differentiate higher proficiency levels: Band 8 – ‘develops topics coherently and appropriately’ and Band 9 – ‘develops topics fully and appropriately’, with little information on the distinctions between ‘coherently’ and ‘fully’. For lower levels, topic is referred to under the Lexical Resource scale to differentiate extent of lexical knowledge on familiar and unfamiliar topics: Band 3 – ‘has insufficient vocabulary for less familiar topics’ and Band 3 – ‘able to talk about familiar topics but can only convey basic meaning on unfamiliar topics’ (see www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en for a public version of the Band Descriptors). What is left unsaid, of course, is how familiarity is determined for different candidates.

A topic or content-oriented criterion capturing the extent to which ideas are developed has not been explicitly defined as part of the construct of IST and other speaking tests more widely (Elder, McNamara, Kim, Pill