# Topic and background knowledge effects on performance in speaking assessment

## Author: Nahal Khabbazbashi

### Abstract

This study explores the extent to which topic and background knowledge of topic affect spoken performance in a high-stakes speaking test. It is argued that evidence of a substantial influence may introduce construct-irrelevant variance and undermine test fairness. Data were collected from 81 non-native speakers of English who performed on 10 topics across three task types. Background knowledge and general language proficiency were measured using self-report questionnaires and C-tests respectively. Score data were analysed using many-facet Rasch measurement and multiple regression. Findings showed that for two of the three task types, the topics used in the study generally exhibited difficulty measures which were statistically distinct. However, the size of the differences in topic difficulties was too small to have a large practical effect on scores. Participants' different levels of background knowledge were shown to have a systematic effect on performance. However, these statistically significant differences also failed to translate into practical significance. Findings hold implications for speaking performance assessment.

### Keywords

Background knowledge, many-facet Rasch measurement, practical significance, speaking performance assessment, topic

In performance-based assessments of speaking, a common practice for eliciting speech samples is to engage test takers with speaking tasks on different topics. In addressing topics, examinees often have to draw on their topic-related background knowledge (BK) defined as 'the information base that enables [individuals] to use language with reference to the world in which they live' (Bachman & Palmer, 1996, p. 65). A facilitative role for BK in second language (L2) performance has been suggested in the theoretical literature where higher levels of BK are associated with lower cognitive demands on individuals by allowing information to become more easily accessible, requiring fewer attentional resources and thus facilitating the retrieval of materials for speech (Skehan, 1998).

In speaking tests where tasks are randomly assigned to candidates, there is an assumption that different topics are of equivalent difficulty. What logically follows is another underlying assumption that the level of BK that individuals bring to the topic does not significantly influence test results. Evidence to the contrary may suggest a potential source of bias if the task includes content that is not equally familiar to different candidate groups (O'Sullivan & Green, 2011) and can introduce a validity threat owing to construct-irrelevant variance (Jennings, Fox, Graves, & Shohamy, 1999). This is of particular concern in *independent* speaking tasks, as opposed to *integrated* tasks, which require candidates to rely on their own ideas and knowledge when responding (Brown, Iwashita & McNamara, 2005) and as such have been criticized for not allowing candidates an 'equal footing' in terms of BK (Weigle, 2004, p. 30).

Concerns about potential topic/BK effect in speaking tests have been voiced by examiners and examinees alike, emphasizing the practical nature of this problem. For example, in a large-scale worldwide survey of 269 IELTS examiners (Brown & Taylor, 2006) topic-related issues emerged as one of the strongest themes. The majority of the

examiners did not find the topics within each part of the test to be equivalent and open comments touched on the level of topic comparability in terms of difficulty, appropriateness and/or complexity as a function of age, culture and language proficiency of candidates.

Candidate concerns with topics on the IELTS speaking test are also documented in Smith's (2009) study. Reflecting on their test day experience, some participants referred to 'overall luck, topic luck and the examiner' as factors potentially affecting their speaking scores. Topic-related problems were repeatedly commented on and included having little to say about the topic (even in their L1), the inability to relate to and/or having little interest in the topic and experiencing anxiety as a result of topic unfamiliarity. Despite these concerns and the increasing use of speaking tests world-wide, there are surprisingly few studies that have systematically examined the effects of topic and BK on speaking. The need for empirical evidence on the comparability of tasks and test forms in large-scale standardized assessment tests such as IELTS has been noted by Weir (2005).

Given (a) the paucity of empirical research, (b) the centrality of topics in creating a meaningful context for eliciting speech, (c) the online nature of speaking which necessitates candidates to draw on their BK for the spontaneous generation of ideas and (d) the real-world concerns of raters and candidates with a potential topic effect on scores, a close examination of these variables on speaking becomes critical from a test validity perspective and constitutes the rationale for this study.

## Literature review

This section provides a review of studies which have examined the role of topic and BK on L2 performance, classified according to the four language skills. The purpose of this review is to establish whether the importance ascribed to topic and BK in the theoretical literature (Bachman & Palmer, 1996; Skehan, 1998) is reflected in empirical research, and to critically examine the methodologies in informing the current study's design.

### *Reading*

Evidence for a lack of a systematic BK effect comes from the seminal work of Clapham (1996) where participants took a reading test relevant to their field of study and one from a different field. BK was established using an *a priori* questionnaire of reading habits and content familiarity. Findings were inconsistent, with some students performing better on tests in their field but only when passages were highly subject-specific. It was also suggested that students may draw on their BK above a certain proficiency threshold. This hypothesis was put to the test by Krekeler (2006) in a study where C-tests and three different BK classification criteria were used as measures of language proficiency and BK respectively. Results suggested significant differences between the reading scores of the two proficiency groups on all BK measures leading to a rejection of the hypothesis. Usó- Juan (2006) also found significant influences of discipline-related BK and proficiency in explaining between 21–31% and 58–68% of the variance in reading scores. Liu (2011), in contrast, failed to find consistent evidence of differential item functioning (DIF) in the reading performances of candidates who were more likely to be favored by specific pas- sages due to their major fields of study and cultural background compared to a reference group. It can be argued that, particularly for pre-entry candidates, major is an indication of interest and as such, a weak proxy for BK.

## *Listening*

Evidence for the positive role of topic familiarity on listening comprehension comes from Schmidt-Rinehart (1994). The study involved students listening to familiar and novel passages. Familiarity was determined *a priori* based on previous exposure to course content while classroom groupings were used as a proficiency measure. Significant effects were found for topic familiarity and proficiency with no significant interaction between the two. A facilitative role for BK was also found in Markham and Latham's (1987) study on the effects of religious-specific BK on listening with a clear trend of increase in mean scores when there was a match between passage content (texts of prayer rituals) and religious background. Jensen and Hansen (1995), on the other hand, failed to find reliable BK effects in understanding academic lectures. The study used binary yes/no passage familiarity questions to classify students into BK and no BK groups. An independent measure of listening proficiency was also included. Multiple regression analyses indicated a significant effect for listening proficiency. However, an effect for BK was found in less than half the lectures with a small effect size, accounting for 3–9% of score variance.

## *Writing*

A positive role for BK was found in Tedick (1990) where the written performance of graduate students was reported to be significantly better – across all proficiency levels on a field-specific topic (which required writers to choose and discuss a controversial issue in their field of study) compared to a general topic. The field-specificity of the prompt was questioned by Lim (2009, p. 37) who found it to be 'ironically… the more general prompt' as it is 'virtually unconstrained, leaving respondents plenty of leeway on what to write about'.

Choice of topic, as an indication of BK, was also used in Jennings et al. (1999) but contrary to Tedick (1990), a comparison of performances of participants randomly assigned to *choice* and *no-choice* conditions revealed no significant differences between the groups. The authors suggested that the integrated nature of the test potentially attenuates the topic effect. Also within an integrated assessment context, Lee and Anderson's (2007) large-scale study examined the impact of academic topics, general proficiency (using standardized TOEFL scores) and BK (operationalized as students' departmental affiliation) on writing. Findings showed a main topic effect once proficiency was controlled for, although no topic × BK interaction was found. This evidence was used to conclude that topics were general enough to be used in the test. However, departmental affiliation is arguably too broad a categorization to be used as a reliable measure of BK. The authors also emphasized the need for independent measures of BK which are more 'critically related' to topics for future studies.

He and Shi (2012) questioned the fairness of including topics which may require cultural or subject-specific knowledge in standardized assessment contexts. In their study, the two prompts of 'university studies' and 'federal politics' were identified as requiring 'general' and 'specific' topical knowledge respectively and were administered to students from three proficiency levels. Findings were illustrative of a strong facilitative role for BK with a clear pattern of significantly higher scores on the general topic. However, these results should be interpreted with some caution, as the use of only two highly oppositional topics in the study's design may have increased the likelihood of observing significant differences in performance.

## *Speaking*

Inconsistent topic effects were reported in Smith's (1989) study with International Teaching Assistants (ITAs) where performances on field-specific vs. general versions of a speaking test were compared across several linguistic features. Results did not reveal a clear pattern with some ITAs performing better on field-specific tests and others on the general version with small, non-significant differences in mean scores. Papajohn's (1999) systematic study compared the spoken performances of prospective ITAs on 15 different chemistry topics, classified on the basis of their cognitive demand. Multiple regression results suggested that general language proficiency and topic classifications were significant predictors of spoken performance accounting for 67.2% and 4.7% of the variance respectively. The only limitation of the study was that rater severity was not accounted for.

Gender-related topic bias was investigated in Lumley and O'Sullivan's (2005) large-scale study in which topics on a tape-mediated speaking test were classified as neutral, male or female-oriented. Analysis using many-facet Rasch measurement (MFRM) suggested instances of topic bias although the bias size was found to be small and not always stable across test forms. Once again, the study's topic classification system based on assumptions of what is stereotypically male or female may have been too simplistic, thus confounding the results.

Bei (2010) explored the influence of task preparedness – operationalized as topic familiarity – on spoken performance of 80 Chinese undergraduates. Similar in design to He and Shi (2012) two parallel tasks were constructed with topics corresponding to the participants' academic disciplines and subsequently administered in matched/mismatched topic familiarity conditions. Results suggested a significant effect for topic familiarity in enhancing fluency, accuracy, lexical sophistication and diversity albeit with a small effect size. Huang (2010) found a strong positive influence of topic knowledge on speaking in one of the only studies which took a rigorous, non-assumption-based approach to measuring BK using a series of extensively piloted topic knowledge tests. While the topic knowledge effect varied across different topics, it was observed in both independent and integrated speaking tasks.

## *Summary and methodological implications*

Research on the effects of topic and BK on performance across the four skills would appear to be mixed and inconclusive. It can be argued that these are largely a result of the following:

(a) a strong reliance on assumption-based indicators of BK as a predictor of task difficulty and on the basis of group level factors such as academic major (Bei, 2010; Lee & Anderson, 2007), gender (Lumley & O'Sullivan, 2005), choice of topic (Jennings et al., 1999; Tedick, 1990), religion (Markham & Latham, 1987) and cultural background (He & Shi, 2012) rather than measuring BK at the level of the individual. These group-level factors may be weak proxies for BK; for example, the use of academic major, particularly for pre-entry candidates, may simply reflect interest in a field of study. Moreover, being from a specific major, cultural background or religion does not necessarily preclude having knowledge about other majors, cultures and religions. Such assumption-based approaches to estimating task difficulty have been problematized by Bachman (2002) who views difficulty as an 'artifact' of test performance and emphasizes the need to clearly

delineate between (i) task-specific features (e.g., topic) which do not require making assumptions about test takers, (ii) test taker characteristics (e.g., BK of topic), (iii) interactions between (i) and (ii) and to subsequently conceptualize and model interactions as such;

(b) study designs that are likely to yield significant results through the use of limited number (often two) of clearly oppositional topics (e.g., He & Shi, 2012; Tedick, 1990);

(c) differences in the way in which language proficiency (as a mediating factor) has been measured in various studies from classroom groupings (Schmidt-Rinehart, 1994) to standardized English tests (Lee & Anderson, 2007);

(d) the complexity of performance assessment contexts in which additional variance can be introduced by a number of different factors including the well-documented rater effect (McNamara, 1996).

The current study addresses these methodological issues by (a) including an independent, non-assumption based measure of BK at the level of the individual; (b) using a range of topics for which test takers are likely to have different levels of BK; (c) using an independent measure of language proficiency; and (d) using MFRM so as to model characteristics of the test task, test taker, rater and any interactions between them systematically, following Bachman (2002).

It should be emphasized that when discussing performance, the focus here is on performance *scores* and not features of *discourse*. Fulcher and Márquez-Reiter (2003, p. 326) direct attention to empirical research which illustrates a lack of '*score sensitivity'* of performance to changes in tasks leading them to challenge the 'unstated assumption that changes in discourse automatically translate into changes in test score'. The importance of establishing '*practical significance*' is then highlighted by Fulcher (2003, p. 65) referring to differences which are not only statistically significant but are also associated with large effect sizes (Kirk, 1996). Dorans and Feigenbaum (1994) advanced the term 'difference that matters (DTM)' in relation to SAT scores where 'any differences less than the DTM are considered not big enough to warrant any concern since they are smaller than the smallest difference that might actually matter' (Dorans & Liu, 2009, p. 13). What constitutes *practically significant* is therefore dependent on specific assessment contexts.

## Research questions

In light of the above discussions, the study's main research question (RQ) is:

> *How is the validity of a speaking test influenced by the random assignment of topics to test takers who bring different levels of BK to the topics?*

The following subsidiary RQs guided the collection of different sources of validity evidence:

(i) To what extent are parallel versions of a speaking test that consist of different topics comparable in terms of difficulty? Are (any) differences large enough to have practical significance?
(ii) When task type is held constant, to what extent are different topics used in parallel versions of a task similar in terms of difficulty? Are (any) differences large enough to have practical significance?
(iii) Do differences in test takers' BK levels have an impact on performance? Are (any) differences large enough to have practical significance?
(iv) Does BK of topics differentially affect performances of test takers from different proficiency levels?

## Method

### *Assessment context*

The assessment context for this study is the IELTS Speaking module, a high-stakes standardized test of English. It is a face-to-face interview between a candidate and

examiner which lasts 11–14 minutes. The construct underlying the test is communicative (spoken) language ability (Seedhouse & Harris, 2011). The test consists of three parts where each part is designed to achieve a particular function in terms of interac- tion pattern, task input and candidate performance. Part 1 (*Information Exchange*) usually consists of two topic sets (or *frames*) where the examiner poses a series of questions on general and familiar topics. In Part 2, the *Individual Long Turn*, the can- didate is required to give an extended monologue for 1–2 minutes on a specified topic. Part 3 (*Two-way Discussion*), typically consists of two topic sets which are more abstract in nature and are thematically linked to the Part 2 topic. The task types corre- sponding to these three parts are henceforth referred to as task types A, B and C. Topic selection and task construction follow standard procedures and item writer guidelines are designed to ensure topic neutrality, minimize bias and increase task comparability (Galaczi & ffrench, 2011). The use of an *Examiner Frame*, 'a script that must be fol- lowed' (IELTS Examiner Training Material 2001, p. 5) ensures reliability of test deliv- ery (Taylor, 2007).

This speaking test largely consists of independent, heavily topic-based tasks with an information-transfer oriented purpose which require candidates to rely on their BK to respond to questions and develop topics. Moreover, candidates are not allowed any choice in topic selection and have little control in topic management owing to the strict examiner frame (Seedhouse & Harris, 2011). Therefore, the effects of topic and BK are arguably of particular salience in this specific test, hence its selection as the study's context.

## *Participants*

Test takers in this study were 81 Farsi speakers of English as a Foreign Language aged between 18 and 40. There were 41 females and 40 males. All were enrolled in IELTS preparation courses in different language centres in Tehran, Iran. They therefore consti- tuted a fairly homogeneous sample in terms of L1, cultural background and exposure to target culture.

The C-test results ($M$ = 0.25; $SD$ = 1.01; range = −2.18 to +3.80 logits) coupled with the speaking test results ($M$ = 5.95; $SD$ = 0.70; range = 3.5 to 8.0 on the IELTS scale) suggest a range of ability levels spanning CEFR levels A2 to C2 (Lim, Geranpayeh, Khalifa, & Buckendahl, 2013).

Four raters (three female, one male; L1 English) participated in the study. They were selected on the basis of their academic qualifications, extensive teaching experience and familiarity with a variety of speaking tests. All raters received training.

## *Design*

A parallel forms reliability design was used where participants responded to two versions of an IELTS speaking test (each consisting of five topics). Participants' BK of topics and general language proficiency were measured using BK questionnaires and C-tests respectively, and their spoken performances were marked by four raters. Resultant data were analyzed using MFRM and multiple regression.

## *Instruments*

*Speaking tasks/tests.* Speaking tasks were selected from a pool of publicly available IELTS materials. It was important to ascertain that, with the exception of task topic, other task-related variables were controlled for so that (any) differences in scores could be predominantly attributed to differences in topics and test takers' BK of topics (Bachman, 2002; Weir, O'Sullivan & Horai, 2006); type of task input was controlled for by following the IELTS speaking test format, the examiner/interlocutor role was fulfilled by a trained IELTS examiner to control for any interlocutor effects (O'Sullivan, 2000), and uniformity of test delivery was ensured by strictly adhering to IELTS administration procedures. A panel of expert colleagues also rated the tasks on a number of different criteria (e.g., lexis, grammar, functions and topic familiarity of the tasks) using a task equivalence checklist (Weir, O'Sullivan & Horai, 2006) and provided open comments.

Final task selection was made on the basis of the panel's ratings, qualitative analysis of their comments and a consideration of task input statistics. This analytic exercise allowed for the selection of task topics which exhibited equivalence to an acceptable degree on a number of different features.

To maximize the number of topics included in the study, an incomplete-connected data collection design (Eckes, 2009; Weir & Wu, 2006) was adopted; four test versions were constructed (W, X, Y, Z) each consisting of two Task A topics, one Task B topic and two Task C topics following the IELTS Speaking Test format (see appendices for an example test version). Two common tasks were used in versions W and Y in order to create the necessary *common link* between the tests, allowing for coverage of 18 different topics (see Table 1).

Participants were divided into two groups; Group 1 responded to versions W and X and Group 2 responded to versions Y and Z resulting in 10 topic-based performances for each participant while ensuring that the requirements of MFRM are met through task overlap. Note that while the two groups were connected through common tasks from only one task type (A), there was full overlap on all task types within each group which ensured construct coverage and supported quality of the equating.

*Rating scale.* The public version of the IELTS Speaking Band Descriptors (IELTS, n.d.) was used for scoring. This nine-band analytic scale consists of four criteria: *Fluency and Coherence* (FC), *Lexical Resource* (LR), *Grammatical Range and Accuracy* (GA) and *Pronunciation* (P). Scores are awarded for each criterion (as whole bands) and subsequently averaged and rounded to the nearest upper half band or whole band. In this assessment context and given the marking model and rounding conventions, a difference of one band on one of the four criteria is considered as the DTM (i.e. the smallest difference that might actually matter to the candidate in terms of their final score).

*Background knowledge questionnaires*. The present study views the interaction between test takers' BK and topic of a task as a complex phenomenon that cannot be assumed or predetermined. BK questionnaires were constructed to capture *relative* degree of topic-related BK and were completed after the speaking tests. The questionnaire consisted of eight questions (Table 2) and responses were elicited on a five-point Likert scale.

**Table 1.** Incomplete-connected data collection design.

| Topics | Task types | Examinee (Group 1) | Examinee (Group 2) |
|---|---|---|---|
| **A.1** | A | X | X |
| **A.2** | A | X | X |
| A.3 | | x | |
| A.4 | A | x | |
| A.5 | | | x |
| | A | | |
| | A | | |
| A.6 | A | | x |
| B.1 | B | x | |
| B.2 | B | x | |
| B.3 | B | | x |
| B.4 | B | | x |
| C.1 | C | x | |
| C.2 | C | x | |
| C.3 | | | |
| C.4 | C | x | |
| C.5 | | | x |
| | C | x | |
| | C | | |
| C.6 | C | | x |
| C.7 | C | | x |
| C.8 | C | | x |

Test Version W: **A.1, A.2**, B.1, C.1, C.2 (Group 1, *N* = 41).
Test Version X: A.3, A.4, B.2, C.3, C.4 (Group 1, *N* = 41).
Test Version Y: **A.1, A.2**, B.3, C.5, C.6 (Group 2, *N* = 40).
Test Version Z: A.5, A.6, B.4, C.7, C.8 (Group 2, *N* = 40).

These questions were *repeated* for *each* topic in order to focus participants' attention on individual topics and elicit any nuances in BK levels. Moreover, care was taken in the phrasing of questionnaire items (e.g., items 1, 3, 4, 5 & 7) to focus on familiarity of topics, availability of ideas, having things to say and interest in topic which were more performance-independent. Both steps were taken to avoid simply eliciting examinees' overall impression of their own performance.

*C-tests.* The literature review identified general language proficiency as one of the variables that can shape the way a test taker's BK interacts with a topic. C-tests were selected as an independent measure of proficiency given their relative ease of development and evidence of high correlations with speaking measures (Eckes & Grotjahn, 2006). C-tests were subsequently constructed, piloted and validated (Khabbazbashi, 2014). Given the integrated nature of language proficiency, a more comprehensive standardized test of English measuring all four skills would have been preferable but was not possible owing to practical constraints. Nevertheless, the steps taken for validating the C-tests ensured that the measurement instrument was reliable and fit for purpose.

## *Procedures*

Participant data collection took place in different language schools during class hours. Participants first completed the C-tests. They were then called out individually to a quiet

**Table 2.** Background knowledge questionnaire items.

1. This topic is familiar to me.
2. The questions about this topic were easy to respond to.
3. I know *A LOT* about this topic, i.e., I have *MORE THAN ENOUGH IDEAS* to talk about this topic.
4. It was easy for me to produce enough ideas for this topic from memory.
5. If I were to talk about this topic in my first language, I would have *MORE IDEAS* to talk about.
6. I had appropriate words to express my ideas about this topic easily.
7. I thought this was an interesting topic.
8. I performed very well on this task.

**Table 3.** The common batch rating design.

| Rater | Common batch (1) | Batch (2) | Batch (3) | Batch (4) | Batch (5) |
|---|---|---|---|---|---|
| Rater 1 | X | x | | | |
| Rater 2 | | | x | | |
| Rater 3 | X | | | x | |
| Rater 4 | | | | | x |
| | X | | | | |
| | X | | | | |

room, where two versions of the speaking test were administered by random rotation and in succession. Participants then completed the BK questionnaires.

Responses to C-tests and BK questionnaires were scored. Each recorded speaking test was edited, divided into its constituent topics and anonymized resulting in 810 speaking files (81 persons × 10 topics). This was done to minimize potential rater halo effect.

Rater training was provided in the form of familiarization meetings and an IELTS standard-setting DVD (UCLES, 2006) which included benchmark performances at different band levels. Following Weir and Wu (2006), an incomplete-connected rating design (see Table 3) was used where speaking files were divided into different batches and each rater was asked to rate a common batch and an additional batch. This allowed for an overlap between raters, meeting the requirements of MFRM. Once ratings were completed, a short interview was conducted with each rater to elicit their views and observations on the performances.

## *Data analysis and preliminary results*

*C-tests.* C-tests were dichotomously scored. Responses to individual items (blanks) within each text were combined to build higher-order polytomous 'super items' and analysed using the Partial Credit derivation of the Rasch model in RUMM 2030 (Andrich, Lyne, Sheridan, & Luo, 2010). This was to address the violation of the assumption of response independence (Marais & Andrich, 2008). The overall fit residual statistics for super-items ($M$ = 0.20; $SD$ = 0.86) and persons ($M$ = −0.18; $SD$ = 0.83) were generally close to their expected values of 0 and 1. The total item trait interaction statistics

($\chi^2 = 31.1$; $p = .83 > .05$) and a high person separation index (PSI) of 0.96 suggested that the required property of invariance was met and that C-tests were able to reliably separate persons from different ability levels. Note that a strong positive correlation ($r = .91$, $p < .001$) was later confirmed between these measures and participants' speaking measures which, taken together, strongly support the use of the study's C-tests as a measure of general language proficiency.

*BK questionnaires.* The BK questionnaire responses were analysed using RUMM 2030. The PSI of 0.90 suggested that, despite its short length, the questionnaire was able to reliably distinguish between persons with different BK levels. Some of the items, however, displayed misfit to the model. Several steps were subsequently taken to improve fit (e.g., identification of disordered thresholds, collapsing of disordered categories, identification of items displaying DIF and using the item-split method (Andrich & Hagquist, 2012) as remedial action). Item 5 nevertheless had to be removed from the analysis due to large misfit. An examination of the Item Characteristic Curves for the remaining items (Figure 1) showed that they met model expectations which, combined with the high reliability indices (PSI = 0.91; $\alpha = 0.92$) provided psychometric support for the use of the instrument. Qualitative evidence for the validity of the questionnaire also came from using the BK measures to extract speech samples where participants had indicated very low or very high BK and cross-checking them with the content of examinees' spoken performance. Results were illustrative of a BK effect (see Khabbazbashi, 2014 for the results of the qualitative study).

*Speaking performances.* The MFRM analysis of speaking scores was carried out using FACETS (Linacre, 2011). A series of four-facet MFRM analyses with examinees, raters, topics and criteria as facets were first run. Resultant topic measurement reports were used to address RQs (i) and (ii). A five-facet MFRM was subsequently run where BK was conceptualized as an additional facet. Given that FACETS only accepts integer numbers, the BK measures were divided into three low, medium and high groups and the resulting BK measurement report was used to address RQ (iii). Participants were then divided into three proficiency levels based on C-test results and an MFRM bias analysis (proficiency x BK) was run to address RQ (iv). Task type, the linear BK and general language proficiency measures were also used as predictor variables for participants' estimated spoken ability measures on each topic in a multiple regression analysis with SPSS.

## Results

The vertical map (Figure 2) visually represents MFRM, displaying the calibrations for all facets in the study. The examinee measurement report illustrated a wide distribution of speaking abilities spanning 8.49 logits. Separation indices suggested a minimum of 12 statistically distinct speaking ability strata ($G = 9.00$, $H = 12.34$) with an associated Rasch person separation reliability of $r = .99$. The four raters in the study exhibited high levels of consistency in their marking with infit statistics falling between stringent lower and upper control limits of 0.7–1.3 (Myford & Wolfe, 2000). The criterion facet results

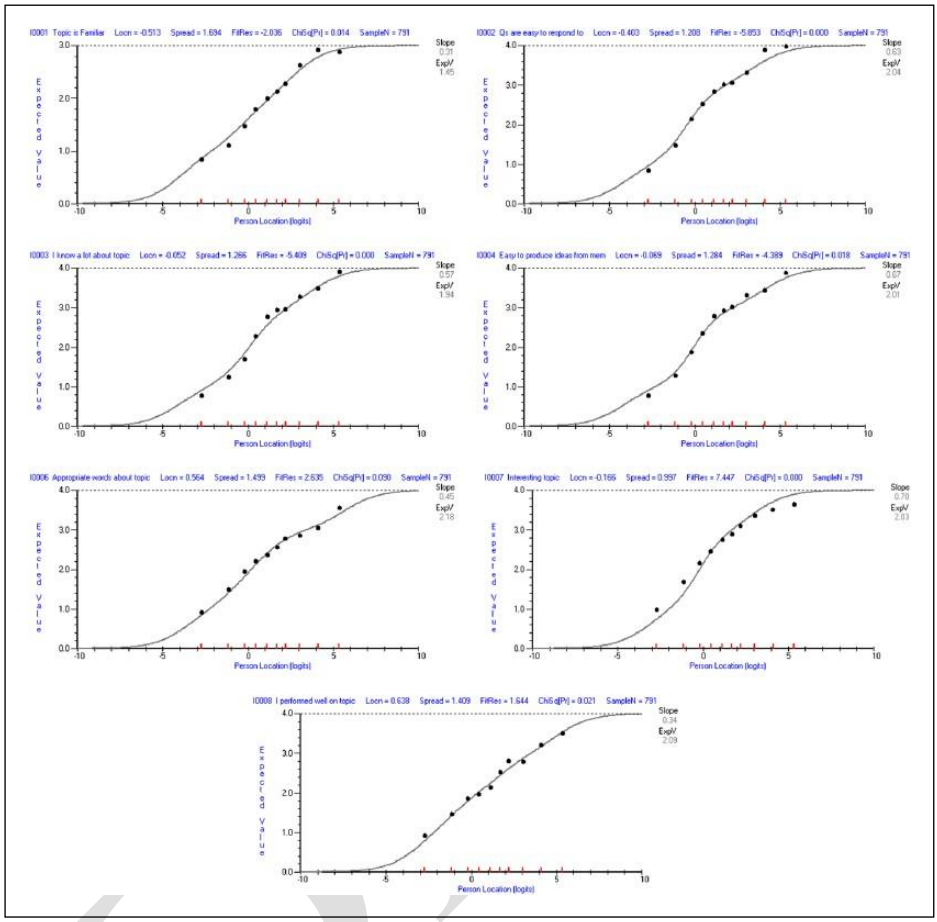


**Figure 1.** Item characteristic curves for questionnaire items 1–4 and 6–8.

suggested that the analytic criteria used in the study (FC, LR, GA and P) contributed in distinct ways to separating examinees into different ability levels. An examination of the rating scale structures and categories showed that the categories within the scales generally functioned well.

## *Topic effects on performance at the test level*

The topic measurement results indicated a relatively narrow range of topic difficulty distribution from the easiest topic (A.6) with a logit value of −0.37 to the most difficult topic (C.4) with a logit value of +0.25, spanning 0.62 logits; the examinee range (8.49) is thus 13.69 times the topic range. The infit and outfit statistics for all 18 topics fell within stringent limits of 0.7 to 1.3. The separation indices (G = 2.21; H = 3.27) suggested that topics could be divided into three statistically distinct difficulty strata with a high

degree of separation, as evidenced in the separation reliability value ($r$ = .83). A significant chi-squared value ($\chi^2 = 104.9$; $df = 17$, $p = .00 < .01$) further rejected the null hypothesis that all topics were of equivalent levels of difficulty. While this was not unexpected given the use of three different task types which are designed to vary in difficulty, the expected clustering of task types from the easiest type (A) to the most difficult (C) is not always observed (Figure 2).

In order to examine the effects of topic at test level, two speaking test versions were constructed: one combining the easiest topics within each task type and one combining the most difficult topics (see Table 4). The average differences between the two versions represented the maximum possible difference attributable to topics, the influence of which was then examined in relation to the average abilities necessary to move across adjacent score categories for the different analytic criteria. This was done by calculating the difference between average abilities observed at adjacent band levels. As discussed earlier, practical significance in the assessment context of the IELTS Speaking Test is operationalized as a difference that would translate into a one-band difference on each of the criteria.

To illustrate, consider the category statistics for the FC criterion in Table 5 where the observed speaking measures at Bands 5 and 6 are −2.50 and −1.40 logits respectively bringing the difference between them to 1.10 logits. This is the average speaking ability measure required to move from Band 5 to 6. For topics to exert a meaningful and practically significant influence on performance, differences attributable to topic difficulties would need to exceed (or be close to) this value.

Following this line of analysis, differences in observed measures across score categories for all criteria were calculated (Table 6). Results in Table 4 showed that the average difficulties for the constructed easy and difficult versions of the speaking test are −0.17 logits and 0.13 logits respectively, bringing the difference between them to 0.31 logits (0.16 IELTS band score), a difference attributed to selected topics. Table 6 shows that the average difference in participants' ability measures at adjacent band scores across criteria is approximately 1.42 logits (ranging between 0.82 and 2.39); a value which consistently and systematically exceeds the maximum topic-related effect ($1.42 > 0.31$).

In answering RQ (i), findings show small and negligible differences in the difficulty of the speaking test versions attributable to differences in topic difficulty. These differences are unlikely to have a significant practical influence on scores within criteria. If the combination of topics from two extreme difficulty levels cannot affect scores at the band level, it can be inferred that, in general, differences in topic difficulties within the assessment context under study are unlikely to have a significant and practical influence on performance at the test level.

## *Topic effects on performance at the task level*

In evaluating the effects of topic at the task level, a series of MFRM analyses were run for each task type. The topic measurement reports for task types A, B and C showed that all topic infit statistics fell within the range of 0.7 to 1.3. The topic separation indices showed that topics could be separated into 1.34 ($r = .36$), 4.24 ($r = .90$) and 3.38 ($r = .84$) difficulty strata for Task Types A, B and C respectively. For Task Type A, these results

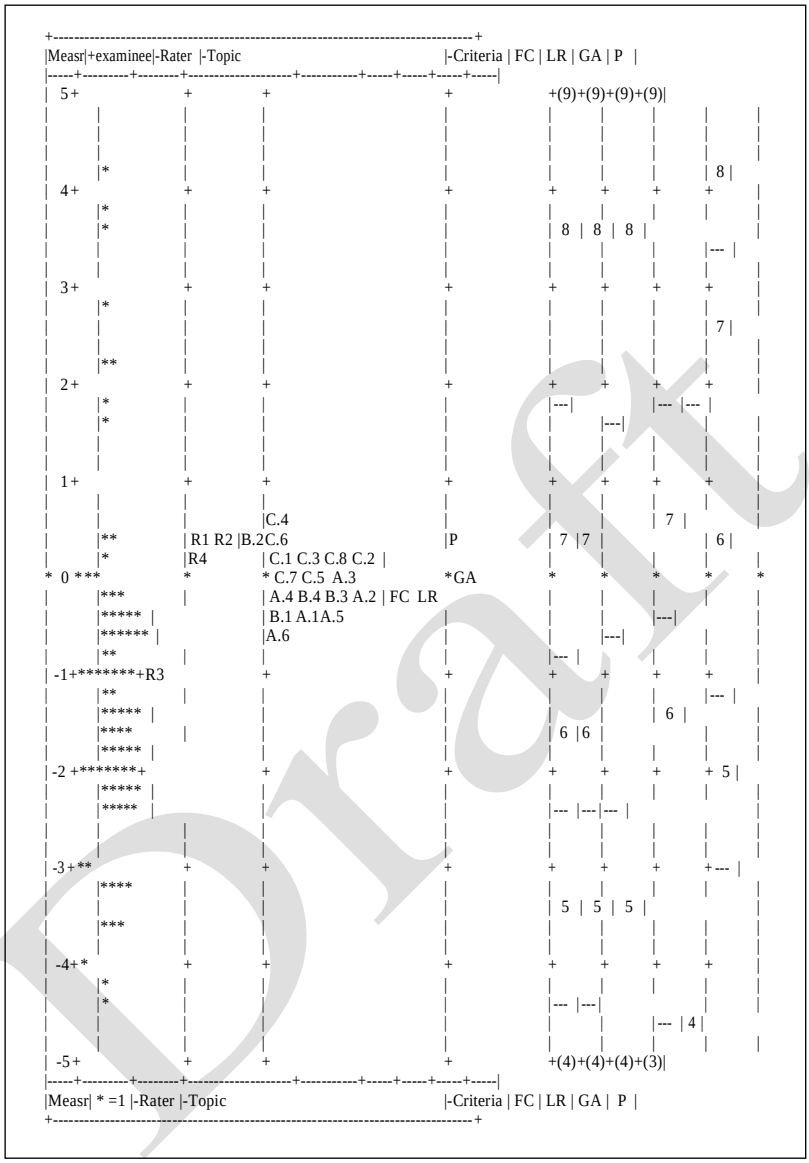


**Figure 2.** Facet map (4-facet MFRM).
*Note*: Each star (*) in the second column represents one examinee.
FC= Fluency and Coherence, LR= Lexical Resource, GA= Grammatical Accuracy, P = Pronunciation.

suggest that the six topics cannot be separated into statistically distinct difficulty levels and are thus considered parallel; a result also substantiated by the non-significant chi-squared statistic. In contrast, for Task Types B and C, the high topic separation indices and associated reliability values indicate that the differences in topic difficulty measures

**Table 4.** Easy vs. difficult versions of the IELTS Speaking Test.

| Test version (easy) | | | | | Test version (difficult) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Topic ID | Task type | Description | Fair–M Avg | Measure (Logits) | Topic ID | Task type | Description | Fair–M Avg | Measure (Logits) |
| A.6 | A | Dancing | 6.21 | −0.37 | A.4 | A | Colour | 6.05 | −0.05 |
| A.5 | A | Keeping in Contact | 6.16 | −0.27 | A.3 | A | Festivals | 6.02 | 0.00 |
| B.1 | B | Friend (Describe) | 6.14 | −0.23 | B.2 | B | River (Describe) | 5.90 | 0.23 |
| C.5 | C | Important Choices | 6.02 | 0.00 | C.6 | C | Choices in Everyday Life | 5.90 | 0.23 |
| C.7 | C | Family Similarities | 6.02 | 0.00 | C.4 | C | Rivers-Economy | 5.89 | 0.25 |
| *Mean (N= 5)* | *Easy test version* | | *6.11* | *–0.17* | *Mean (N= 5)* | *Difficult test version* | | *5.95* | *0.13* |

Difference in average measures of the two versions = 0.13 – (−0.17)= 0.31 (logits).
Differences in Fair–M Average Measures = 6.11–5.95 = 0.16 IELTS band score.

**Table 5.** Category statistics: Fluency and coherence.

| Band | Counts | Avg measure | Outfit mean square |
|---|---|---|---|
| 4 | 83 | −3.54 | 0.8 |
| 5 | 257 | −2.50 | 0.8 |
| 6 | 446 | −1.40 | 0.8 |
| 7 | 438 | −0.27 | 0.9 |
| 8 | 165 | 1.96 | 1.0 |
| 9 | 13 | 3.68 | 1.3 |

**Table 6.** Increase in average ability measures in adjacent categories (test level).

| Bands | FC | LR | GA | P |
|---|---|---|---|---|
| 3–4 | | | | 0.82 |
| 4–5 | 1.04 | 1.08 | 1.16 | 1.10 |
| 5–6 | 1.10 | 1.11 | 1.19 | 1.03 |
| 6–7 | 1.13 | 1.17 | 1.19 | 2.39 |
| 7–8 | 2.23 | 1.94 | 2.02 | 1.34 |
| 8–9 | 1.72 | 1.81 | 1.54 | 1.18 |
| *Average* | *1.44* | *1.42* | *1.42* | *1.41* |

FC= Fluency and Coherence, LR= Lexical Resource, GA= Grammatical Accuracy, P = Pronunciation.

within each task type are statistically significant and that the topics cannot be considered parallel, a result also substantiated in the significant chi-squared values.

Table 7 summarizes the information extracted from category statistics for each criterion and for the different task types, where differences between average ability measures at adjacent band scores were calculated for all criteria. For each task type (and at the bottom of the section), the maximum difference between the easiest and most difficult topics is also included both in logits and in terms of the IELTS scale, calculated from the topic measurement reports. For example, in Task Type A, this maximum difference is calculated as 0.24 logits (0.12 IELTS bands). For the LR criterion in Task Type A, the lowest average ability required to move across two adjacent band scores is 1.23 logits (Bands 5–6). Given that 1.23 > 0.24, it is unlikely for the differences in topic difficulty measures in Task Type A to have a meaningful influence on performance in the LR criterion at the task level.

When applying the same analysis for different criteria and across the task types, similar results emerge; the maximum difference between the easiest and most difficult topics at each task type are 0.24, 0.52 and 0.54 (corresponding to 0.12, 0.27 and 0.27 on the IELTS raw-score metric) for Task Types A, B and C respectively, none of which exceed the minimum average ability required to move along adjacent score categories for the different criteria. The likelihood of a meaningful topic influence on scores is therefore minimal; a scan of the data in each cell of the table confirms that this is the case, as the speaking ability required to move along adjacent band scores for all the IELTS criteria

**Table 7.** Increase in average ability measures in adjacent categories (task level).

| | Bands | FC | LR | GA | P |
|---|---|---|---|---|---|
| Task type A | | | | | |
| | 3–4 | | | | 1.38 |
| | 4–5 | 1.31 | 1.42 | 1.44 | 1.14 |
| | 5–6 | 1.08 | 1.23 | 1.22 | 1.04 |
| | 6–7 | 1.25 | 1.31 | 1.42 | 2.26 |
| | 7–8 | 2.24 | 1.74 | 1.88 | 1.10 |
| | 8–9 | 1.63 | 1.63 | 1.49 | 1.19 |
| | Average | *1.50* | *1.47* | *1.49* | *1.35* |
| | Task Type A Topics (Maximum Difference in Difficulty) = 0.24 logits/ 0.12 IELTS Bands | | | | |
| Task type B | | | | | |
| | 3–4 | | | | |
| | 4–5 | 1.27 | 1.31 | 1.57 | 1.45 |
| | 5–6 | 1.07 | 1.19 | 1.25 | 0.83 |
| | 6–7 | 1.39 | 1.18 | 1.20 | 2.28 |
| | 7–8 | 2.35 | 2.27 | 2.04 | 1.85 |
| | 8–9 | 2.04 | 2.11 | 2.08 | |
| | Average | *1.62* | *1.61* | *1.63* | *1.33* |
| | Task Type B Topics (Maximum Difference in Difficulty) = 0.52 logits/ 0.27 IELTS Bands | | | | |
| Task Type C | | | | | |
| | 3–4 | | | | 1.17 |
| | 4–5 | 1.00 | 1.09 | 1.15 | 1.16 |
| | 5–6 | 1.27 | 1.23 | 1.22 | 1.26 |
| | 6–7 | 1.20 | 1.25 | 1.38 | 2.66 |
| | 7–8 | 2.37 | 2.34 | 2.24 | 1.71 |
| | 8–9 | 1.81 | 1.46 | 1.45 | 0.88 |
| | Average | 1.53 | 1.47 | 1.49 | 1.47 |
| | Task Type C Topics (Maximum Difference in Difficulty) = 0.54 logits/ 0.27 IELTS Bands | | | | |

FC= Fluency and Coherence, LR= Lexical Resource, GA= Grammatical Accuracy, P = Pronunciation.

consistently exceeds the maximum difference between the easiest and most difficult top- ics for each task type.

In answering RQ (ii), topic separation indices suggest that when task type is held constant, the topics in task types B and C are significantly different and can be divided into a minimum of two difficulty strata; that is, there are at least two topics within each task type where differences in difficulty measures are statistically significant, whereas the six Task Type A topics exhibited very similar difficulty measures and could therefore be considered parallel. However, even in Task Types B and C where at least two of the topics belong to statistically distinct difficulty strata, the differences were not large enough to be translated into meaningful differences in performance scores, thus failing to exert practical significance.

## *BK effects on performance*

The BK estimates for each person × topic combination were first divided into three groups (Low, Medium and High) and a new MFRM analysis was run with BK as an additional facet. The literature review suggested a potentially facilitative effect of higher levels of BK on performance. On this basis, we would assume the Low BK condition to be the most challenging condition (higher logit value) and the High BK condition to be the easiest condition (lower logit value). On the other hand, if BK does not exert an influence on performance then the different BK conditions would not appear in any particular order and their measures would be very close in difficulty.

A consideration of the five-facet vertical map (Figure 3) suggests that not only are the different BK conditions ordered as predicted, that is, from high to low in ascending order of difficulty, but that there is also a notable distance between the BK element measures. This preliminary observation supports the facilitative effect of higher-level BK on performance.

The BK measurement report (Table 8) shows that the High BK condition is associated with the lowest measure (High BK = −0.29) whereas the Low BK condition has the highest measure (Low BK = +0.34) spanning a range of 0.63 logits (0.32 on the IELTS raw-score metric). The infit mean square statistics fall between 0.92 and 1.10 and are very close to their expected value of 1.0.

The interpretation of the separation indices for BK is similar to that of raters; just as it is not desirable for relative severity of raters to introduce measurement error to an assessment context, BK should also not exert a significant influence on performance. Ideally, BK separation indices should be low and the separation reliability value close to 0. However, the separation indices ($G = 9.33$; $H = 12.78$; $r = .99$) suggest that BK conditions can be reliably separated into approximately 12 statistically distinct difficulty strata with a high degree of separation between levels as evidenced in the high reliability value of .99. The null hypothesis that these measures are the same is rejected based on the significant chi-squared statistics ($\chi^2 = 169.6$; $p = .00 < .01$). These results confirm that BK can have a statistically significant impact on performance.

The extent to which this influence has practical significance was examined next by looking at the BK effect in relation to average examinee ability levels necessary to move across score categories in the different criteria. Table 9 reveals that the minimum average spoken ability required to move along adjacent band levels (approximately 1.45 logits) consistently exceeds the maximum difference between the lowest and highest BK condition measures ($1.45 > 0.63$), making it unlikely for BK to have a practical effect on scores.

In answering RQ (iii), these findings suggest that while different levels of BK can pose significantly distinct levels of challenge for test takers, these differences do not translate into practical significance.

The MFRM analysis with BK as a facet has two limitations: first, the incomplete design of the study meant missing BK data, as not all participants had BK measures associated with all topics. Secondly, the BK measures could not be directly used in the analyses and had to be grouped into levels, as FACETS only accepts integer numbers. To address these problems, a different statistical technique – multiple regression – was used to examine the relative contribution of the three variables of general language proficiency, BK and task type in predicting spoken performance. To run the analysis, the data was rearranged to treat

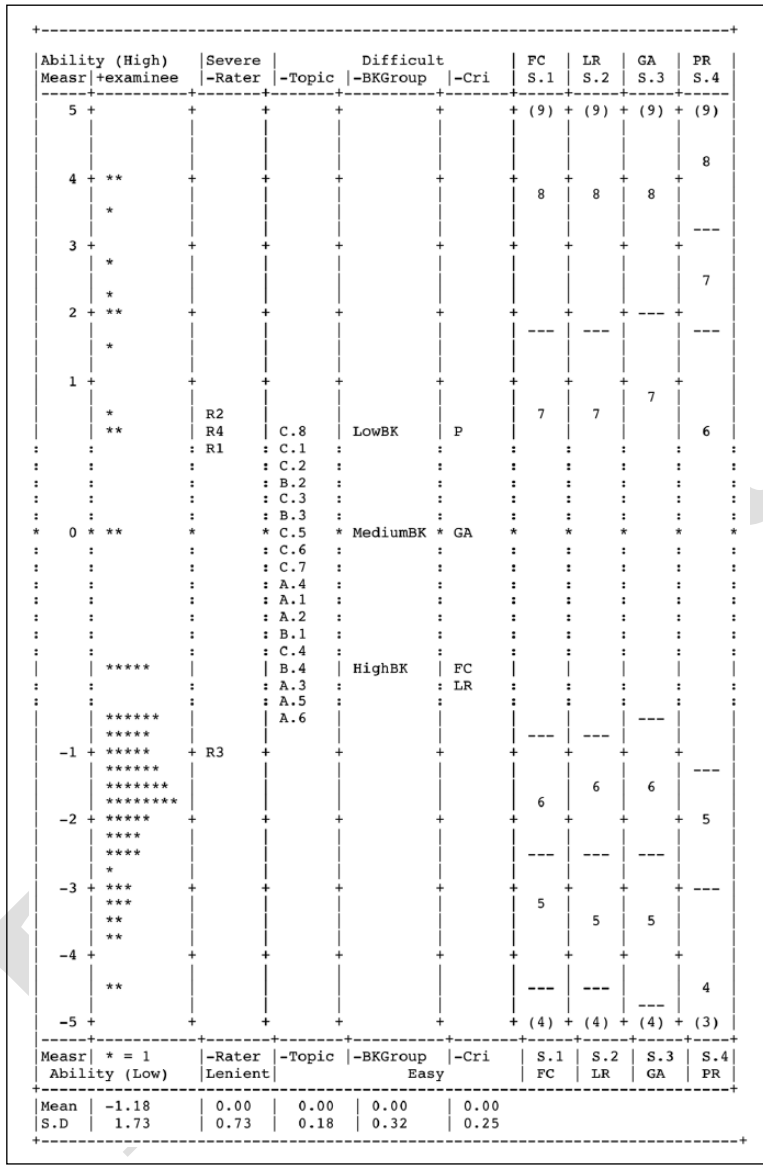


**Figure 3.** Facet map (5-facet MFRM).
*Note*: Each star (*) in the second column represents 1 examinee.
FC= Fluency and Coherence, LR= Lexical Resource, GA= Grammatical Accuracy, P = Pronunciation.

each person × topic combination as a distinct person and a three-facet MFRM analysis (examinee, rater and criteria as facets) was run to estimate speaking measures for the 10 different topics. In this approach, topic was no longer considered a separate facet; instead, relative topic difficulties were absorbed in the resulting person x topic speaking measures but adjusted for any differences in rater severity and task type difficulty.

**Table 8.** The background knowledge measurement report.

| BK level | Obs avg | Fair-M avg | Measure | Model *S.E.* | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| High BK | 6.20 | 6.17 | −0.29 | 0.04 | 1.10 | 1.12 |
| Medium BK | 5.90 | 6.05 | −0.05 | 0.03 | 0.92 | 0.92 |
| Low BK | 5.90 | 5.85 | 0.34 | 0.03 | 0.95 | 0.97 |
| *Mean (N= 3)* | *6.0* | *6.03* | *0.00* | *0.03* | *0.99* | *1.00* |
| *SD* | *0.20* | *0.16* | *0.32* | *0.00* | *0.09* | *0.10* |

Model, Sample: RMSE .03 Adj (True) *SD* .32 Separation 9.33 Strata 12.78 Reliability .99. Model, Fixed (all same) chi-square: 169.6 *df*: 2 significance (probability): .00.

**Table 9.** Increase in average ability measures in adjacent categories (test level).

| Bands | FC | LR | GA | P |
|---|---|---|---|---|
| 3–4 | | | | 0.89 |
| 4–5 | 1.13 | 1.14 | 1.22 | 1.13 |
| 5–6 | 1.11 | 1.17 | 1.21 | 1.05 |
| 6–7 | 1.18 | 1.18 | 1.23 | 2.40 |
| 7–8 | 2.25 | 1.98 | 2.06 | 1.38 |
| 8–9 | 1.78 | 1.76 | 1.52 | 1.20 |
| *Average* | *1.49* | *1.45* | *1.45* | *1.43* |

FC= Fluency and Coherence, LR= Lexical Resource, GA= Grammatical Accuracy, P = Pronunciation.

Where order of causality is concerned, the analysis showed a number of cases where estimated speaking measures on specific topics did not necessarily correspond to participants' BK measures. For example, participant XX01 (see Table 10) had a much higher speaking measure on the topic he reported to have lower BK of compared to another topic where higher BK was reported but with a lower corresponding speaking measure. This serves as further evidence that the BK questionnaire was not simply reflecting examinees' perceptions of their performance on the test.

This approach not only allowed for different Rasch-calibrated measures to be directly used in the analysis but also allowed for a comparison of results against other studies in the literature which have used similar techniques. The three predictor variables were entered in the multiple regression using a hierarchical method of data entry. Results showed that general language proficiency and BK accounted for approximately 60% and 1% of variation in scores respectively; F-ratios for both models were significant at the .001 level. Task type, however, was not a significant predictor of performance. These findings are in line with the MFRM results and can be used to answer RQ (iii); while BK is a significant predictor of spoken performance, the effect size is small with minimal impact on scores.

## *BK interaction with proficiency*

In order to examine whether BK differentially affects examinees from different proficiency levels, a BK × proficiency bias analysis was run in FACETS where BK and C-test measures were divided into three Low, Medium and High groups. Findings showed two significant bias terms with *z*-score values larger than |2| (see Table 11). First, a positive measure

**Table 10.** Illustrative example of inverse relationship between BK self-reports and speaking measures.

| Person ref | Topic | BK measure | Speaking measure (on Topic) |
|---|---|---|---|
| XX01 | C.4 | 5.45 (H) | 3.01 |
| XX01 | C.8 | −3.56 (L) | 4.17 |

**Table 11.** Summary of bias analysis results (background knowledge × proficiency).

| Raw average (obs-exp) | Bias measure | Bias model S.E. | Bias Z-score | Bias infit MnSq | Bias outfit MnSq | BK group | Proficiency level |
|---|---|---|---|---|---|---|---|
| 0.06 | 0.14 | 0.07 | 2.05 | 1.20 | 1.30 | Low | High |
| −0.06 | −0.10 | 0.07 | −2.15 | 1.00 | 1.00 | Low | Low |

of 0.14 was observed for persons from a Low BK but High Proficiency grouping which can be interpreted as follows: when BK levels are low, higher proficiency participants are at an advantage compared to participants with low or medium proficiency levels. The second negative bias of −0.10 suggests that persons with Low Proficiency and Low BK are at a disadvantage compared to persons from medium or high proficiency levels.

In answering the final RQ (iv), results suggest that low levels of BK differentially affect persons with high and low proficiency levels, placing the latter at an advantage but disadvantaging the former. There was no evidence of an opposite trend, that is, high BK was not shown to favour or be biased against persons from different ability levels. Note that despite reaching statistical significance, the bias measures were very small (0.1 logits) and therefore unlikely to exert a large effect on outcomes.

## Discussion

The study's main research question delineated topic and BK as factors which can have an impact on performance and potentially introduce construct-irrelevant variance into speaking performance assessment. Previous studies which had attempted this question had done so indirectly and often through assumption-based proxies which did not do justice to the intricate ways in which examinees' individual BK can interact with task topics.

In this study topics generally exhibited difficulty measures which were statistically distinct, but these differences were not large enough to have a practical impact on scores at the test or task level, results that strongly resonate with Fulcher's (2003, p. 64) argument that changes in task conditions do not '*automatically translate into changes in test score*' (*italics* in original).

Results also suggested a statistically significant effect for BK, where low levels of BK posed the greatest level of challenge for test takers while high levels of BK were shown to have a facilitative effect on performance. While it might be possible to interpret this as causality in a different order – where BK responses are simply illustrative of examinees' impression of their performance – various steps were taken to ensure that there is no such ambiguity. Moreover, these findings are in line with the results of various studies which also found a statistically significant role for BK in L2 performance (e.g., He & Shi, 2012;

Huang, 2010; Krekeler, 2006; Schmidt-Rinehart, 1994; Tedick, 1990). Despite reaching statistical significance, BK did not have a practical effect on scores, as the minimum average spoken ability required to move along adjacent band levels was shown to consistently exceed the maximum difference between the lowest and highest BK conditions. BK accounted for only 1% of the variance in scores while general proficiency accounted for 60%. This relatively small BK effect is also in line with empirical findings from other studies; 3-9% of variance in listening scores was attributed to prior knowledge in Jensen and Hansen (1995) and Papajohn (1999) who found topic groupings and general language proficiency to account for 4.7% and 67.2% of variance in speaking scores respectively. In the words of Skehan, Xiaoyue, Qian and Wang (2012, p. 178), it appears that 'speaking about something one is familiar with does produce performance advantages in… various measures, but the advantage is surprisingly small'. While findings indicate the presence of some limited topic-related bias in the test, the observed bias was not large enough to 'distort' test results dramatically or contest the plausibility of interpretations on the basis of test scores (Kane, 2001; McNamara & Roever, 2006).

A possible explanation for the observed small BK effect is the multi-question format of the speaking tasks as commented by one of the raters in the study:

> It's clear to me that the test can get away with questions like 'where can you get information about genetic research in your country?' because even if candidates plead ignorance and know absolutely nothing about it, there are follow-up questions that can bail them out.

In other words, while BK might exert a strong influence at the question level within a topic sequence, the multi-question format of the task serves as a control mechanism for reducing the impact of lack of BK, as speech is likely to be generated by other questions on the same topic. By extension, the multi-topic format of the speaking test further safeguards against the negative impact of BK on scores, as the availability of a minimum of five topics at the test level ultimately reduces the likelihood of *all* topics being unfamiliar.

However, this does not always mean that lower levels of BK do not exert a negative affective influence on candidates (Jennings et al., 1999) or that they do not impact the language intended to be elicited by the given question/task (Seedhouse & Harris, 2011). The rater comments in this study suggested that when encountering an unfamiliar topic, some individuals appeared to 'be caught off-guard', feel 'baffled', 'surprised' or at times, 'scared' and that candidates could not really use complex grammar or lexis when saying that they know nothing about a topic[1]. Steps should therefore be taken to minimize (any) negative impact of topics and BK to reduce test taker anxiety, ensure that candidates perform to the best of their ability and that tasks generate sufficient samples of speech and elicit the type of language intended by task designers. Suggestions may include the implementation of a choice mechanism which reduces the probability of conflict in terms of topic mismatch with candidate BK or the inclusion of some speaking tasks which do not require examinee reliance on BK.

## Concluding remarks

The present study has contributed to an understanding of the role of topic and BK of topics in L2 performance assessment. The study employed a rigorous methodology to

examine the effects of the variables of interest and systematically controlled for the influence of factors which may have confounded results. As such, the findings provide empirical evidence for addressing important validity concerns in speaking tests.


## Acknowledgements

This article is based on my DPhil thesis and I would like to express my deepest gratitude to my mentor and supervisor, Dr Robert Vanderplank, for his guidance and encouragement. I am especially grateful to my friend and colleague, Dr Gad Lim, for patiently reading and commenting on multiple drafts of the manuscript and for his continuous support. My sincere thanks to Professor David Andrich for teaching me about Rasch analysis and to Dr Evelina Galaczi, Professor Ernesto Macaro and Professor Barry O'Sullivan for their constructive feedback. Lastly, I would like to thank the journal editor and the anonymous reviewers for their insightful comments on an earlier draft of this article.


## Note

1. Owing to space limitations, the qualitative findings of the study could not be discussed in detail here.

## References


Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*, 387–416.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2010). *RUMM 2030*. Perth: RUMM Laboratory.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476.

Bachman, L. F., & Palmer, A. S. (1996).*Language testing in practice*. Oxford: Oxford University Press.

Bei, X. (2010).The effects of topic familiarity and strategic planning in topic-based task performance at different proficiency levels. (Unpublished PhD thesis). Chinese University of Hong Kong, China.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

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.

Brown, A., & Taylor, L. (2006). A worldwide survey of examiners' views and experience of the revised IELTS Speaking test. *Cambridge ESOL Research Notes*, *26*, 14–18.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Studies in Language Testing 4. Cambridge: Cambridge University Press.

Dorans, N., & Feigenbaum, M. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. *Technical issues related to the introduction of the new SAT and PSAT/ NMSQT*, 91–122.

Dorans, N., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations*. ETS Research Report No.RR-09–08. Princeton, NJ: Educational Testing Service.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe: Language Policy Division. Retrieved from www.coe.int/t/dg4/Linguistic/ CEF-refSupp-SectionH.pdf.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.

Fulcher, G., & Márquez-Reiter, R. (2003) Task difficulty in speaking tests. *Language Testing*, *20*(3), pp. 321–344.

Galaczi, E., & ffrench, A. (2011). Context validity. In L. Taylor (ed.), *Examining speaking* (pp. 112–170). Cambridge: Cambridge University Press.

He, L., & Shi, L. (2012).Topical knowledge and ESL writing. *Language Testing*, *29*(3), 443–464.

Huang, H. T. D. (2010). Modeling the relationships among topical knowledge, anxiety, and integrated speaking test performance: A structural equation modeling approach. (Unpublished PhD. thesis). University of Texas.

International English Language Testing System (IELTS) (n.d.). *What is IELTS?* IELTS. [Online]. Available at: www.ielts.org/pdf/Speaking%20Band%20descriptors.pdf (Accessed 15 December, 2012).

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing*, *16*(4), 426–456.

Jensen, C., & Hansen, C. (1995).The effect of prior knowledge on EAP listening-test performance. *Language Testing*, *12*(1), 99–119.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Khabbazbashi, N. (2014). An investigation into the effects of topic and background knowledge of topic on second language speaking performance assessment in language proficiency interviews. (Unpublished PhD thesis). University of Oxford.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing*, *23*(1), 99–130.

Lee, H. K., & Anderson, C. (2007).Validity and topic generality of a writing performance test. *Language Testing*, *24*(3), 307–330.

Lim, G. S. (2009). Prompt and rater effects in second language writing performance assessment. (Unpublished PhD thesis). University of Michigan.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, *13*(1), 32–49.

Liu, O. L. (2011). Do major field of study and cultural familiarity affect TOEFL® iBT reading performance? A confirmatory approach to differential item functioning. *Applied Measurement in Education*, *24*(3),235–255.

Linacre, J. M. (2011). *Facets computer program for many-facet Rasch measurement*. Beaverton, Oregon: Winsteps.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
Lumley, T., & O'Sullivan, B. (2005).The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415–437.
Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.
Markham, P., & Latham, M. (1987).The influence of religion-specific background knowledge on the listening comprehension of adult second-language students. *Language Learning*, *37*(2), 157-–170.
McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.
McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.
Myford, C. M., & Wolfe, E. W. (2000) *Monitoring sources of variability within the test of spoken English assessment system (Research Report)*. Princeton, NJ: Educational Testing Service.
O'Sullivan, B. (2000). Towards a model of performance in oral language testing. (Unpublished PhD thesis). University of Reading, UK.
O'Sullivan, B., & Green, A. (2011). Test taker characteristics. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 36–64). Studies in Language Testing 30. Cambridge: UCLES/Cambridge University Press.
Papajohn, D. (1999). The effect of topic variation in performance testing: The case of the chemistry TEACH test for international teaching assistants. *Language Testing*, *16*(1), 52–81.
Schmidt-Rinehart, B. C. (1994). The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal*, *78*(2), 179–189.
Seedhouse, P., & Harris, A. (2011).Topic development in the IELTS Speaking Test. In J. Osborne (Ed.), *IELTS Research Reports*, Vol. 12 (pp. 69–124).IDP IELTS Australia and British Council.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
Skehan, P., Xiaoyue, B., Qian, L., & Wang, Z. (2012). The task is not enough: Processing approaches to task-based performance. *Language Teaching Research*, *16*(2), 170–187.
Smith, J. (1989). Topic and variation in ITA oral proficiency: SPEAK and field-specific tests. *English for Specific Purposes*, *8*, 155–167.
Smith, S. (2009) *IELTS Examination Preparation among University of Oxford Post-graduate students*. (Unpublished MSc thesis). University of Oxford.
Taylor, L. (2007). The impact of the joint-funded research studies on the IELTS Speaking Module. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment* (pp. 185–194). Studies in Language Testing 19. Cambridge: Cambridge University Press.
Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*(2), 123–143.
University of Cambridge ESOL Examinations. (2006). *IELTS Scores Explained (DVD)*.
Usó-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English-language proficiency in reading English for Academic Purposes. *The Modern Language Journal*, *90*(2), 210–227.
Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(1), 27–55.
Weir, C. J. (2005) *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
Weir, C. J., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in speaking tasks: An intra-task perspective. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports*, Vol. 6 (pp. 119–160). United Kingdom, Australia: IELTS Australia and British Council.
Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167–197.

# Appendix

## *Speaking Test (Version Z)*

*PART 1 (Task Type A)*

**Topic A.5: Keeping in contact with people**

**Let's talk about keeping in contact with people.**

- How do you usually contact your friends? [Why?]
- Do you prefer to contact different people in different ways? [Why?]
- Do you find it easy to keep in contact with friends and family? [Why/Why not?]
- In your country, did people in the past keep in contact in the same way as they do today? [Why/Why not?]

(UCLES, 2009: 32)

**Topic A.6: Dancing**

**Now let's move on to talk about dancing.**

- Do you enjoy dancing? [Why/ Why not?]
- Has anyone ever taught you to dance? [Why/Why not?]
- Tell me about any traditional dancing in your country.
- Do you think that traditional dancing will be popular in the future? [Why/Why not?] (UCLES, 2008: 32)

*PART 2 (Task Type B)*

**Topic B.4 : Describe someone in your family who you like.**

Describe someone in your family who you like.

You should say:

How this person is related to you
What this person looks like What
kind of a person he/she is

And explain why you like this person. (UCLES, 2008: 32)

*PART 3 (Task Type C)*

**Topic C.7: Family Similarities**

- In what ways can people in a family be similar?
- Do you think that daughters are always more similar to mothers than to male relatives? What about sons and fathers?
- In terms of personality, are people more influenced by their family or their friends? In what ways?

**Topic C.8: Genetic Research**

- Where can people in your country get info about genetic research?
- How do people in your country feel about genetic research?
- Should this research be funded by governments or private companies? Why? (UCLES, 2008: 32)