Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327

Arab World English Journal (AWEJ) Volume 14. Number 4 December 2023 Pp.213-232
DOI: https://dx.doi.org/10.24093/awej/vol14no4.13

Topic Familiarity Effects on Performance in Speaking Assessment Tasks

Nurul Iman Ahmad Bukhari
Faculty of Language Studies and Human Development, Universiti Malaysia Kelantan, Kelantan, Malaysia
Faculty of Educational Studies, Universiti Putra Malaysia, Selangor, Malaysia

Lilliati Ismail
Faculty of Educational Studies, Universiti Putra Malaysia, Selangor, Malaysia
Corresponding Author: lilliati@upm.edu.my

Noor Lide Abu Kassim
Kuliyyah of Education, International Islamic University Malaysia, Kuala Lumpur, Malaysia
Corresponding Author: noorlide@iium.edu.my

Abu Bakar Razali
Faculty of Educational Studies, Universiti Putra Malaysia, Selangor, Malaysia

Nooreen Noordin
Faculty of Educational Studies, Universiti Putra Malaysia, Selangor, Malaysia

Muhamad Firdaus Mohd Noh
Sekolah Rendah Agama Bersepadu Segamat, Johor, Malaysia

Received:07/11/2023 Accepted: 11/28/2023 Published: 12/15/2023

Abstract

Speaking assessment is believed to be difficult in its expansion and execution. Thus, it is a challenge for teachers in preparing students for speaking tests. This study's purpose was to identify whether topic familiarity affects speaking performance among students who were preparing to sit for the high-stakes Malaysian University English Test speaking test. The study aimed to investigate the validity and reliability of the speaking assessment measures and, subsequently, to examine the speaking tests' item difficulty measures differences according to topic familiarity level. Data were collected from 40 non-native speakers of English among Malaysian Form Six pre-university students who were preparing for their MUET test. The researcher conducted two practice speaking tests, which used retired papers of the MUET CEFR-aligned speaking tests, to the 40 participants who were grouped into 10 speaking test groups. The practice speaking tests were video and audio-recorded. Topic familiarity was measured using self-report questionnaires. In the second phase, each of the seven appointed raters scored all 40 students' responses in two speaking tests consisting of two speaking tasks assessed across four criteria: task fulfillment, language, communicative ability, and group discussion. Many-facet Rasch measurement results revealed significant differences translating to significant influence of topic familiarity on speaking performance. The present study's results not only confirmed the significance of topic familiarity in the preparation for speaking assessments, but also highlighted the need for formal teaching on potential topic themes that are commonly encountered in such assessments, particularly those that are at the post-secondary level. These findings imply importance in designing speaking tests taking into account test-taker topic familiarity.

Keywords: language testing, many-facet Rasch measurement, oral performance, speaking assessment, topic familiarity

Cite as: Bukhari, N. I.A., Ismail, L., Abu Kassim, N.L., Abu Bakar Razali, A., Noordin, N., & Noh, M. F. M. (2023). Topic Familiarity Effects on Performance in Speaking Assessment Tasks. *Arab World English Journal, 14* (4) 213-232. DOI: https://dx.doi.org/10.24093/awej/vol14no4.13

Introduction

In English language testing, it is crucial to understand the factors that influence task difficulty so test tasks can be tailored to best suit target test-takers, such that neither the extremes of difficulty nor leniency are unreasonable. Test designers will profit from this by being able to create tests with higher levels of validity and fairness, which will lead to test-takers being able to be given more accurate decisions, particularly those who are sitting for high-stakes tests.

Moreover, initiatives have been attempted to emphasise the task-based aspect of college-level ESL instruction (e.g., Ismail & Abd. Samad, 2014). The Malaysian University English Test (MUET) format was revised by matching its test specifications with CEFR descriptors because the CEFR is task-based (Council of Europe, 2001; Fischer, 2020). In order to assess secondary school leavers' proficiency in the English language, the MUET was developed in 1999. The test evaluates a candidate's listening, reading, writing, and speaking skills in English. In 2021, the MUET was modified into a proficiency test that was aligned with the CEFR with the assistance of the English Roadmap 2015–2025. This demonstrates how crucial it is for the nation to increase its citizens' proficiency in English-language communication (Chonghui, 2019). This proficiency test is required for some employment in Malaysia and is utilised by the majority of higher education institutions to position students in the correct English ability levels related to their respective programmes (Rethinasamy & Chuah, 2011). Because the majority of MUET candidates use the test to gain admission to higher education institutions, test-takers have a lot at stake. Therefore, the MUET has undergone CEFR alignment as well as rigourous validation procedures conducted by the Malaysian Examination Council (MEC) together with Cambridge Assessment English (Geranpayeh & Ahmad Zufrie, 2018; Malaysian Examinations Council, 2019). Despite this, the MUET speaking test is nevertheless vulnerable to concerns of variability, much like other tests of speaking proficiency have been shown to be. There have been a plethora of studies conducted on the issue of variability, and many factors have been identified and segmented into different categories. Variability in rater-mediated exams, such as the speaking test, may be due to rater behaviour issues (Aryadoust et al., 2021; Bijani, 2018, 2019; Khabbazbashi, 2017) as well as task characteristics. On the other hand, sources of variability might also be garnered from personal traits, more specifically, test-taker characteristics (Abu Kassim & Zubairi, 2006; Bachman, 1999). Hence, there is a need to fulfill the gap in the literature and delve into test-taker characteristics of Malaysian test-takers, which could be a cause of the variability that occurs in the speaking test performance in the MUET test task.

Zooming into test-taker characteristics, whether experiential or psychological, are among the variable contribution which Skehan (2014) claims influences task difficulty, and subsequently influences task performance. For the speaking test, several “parallel” versions with different topics are made to make sure it is fair. This means that a test-taker (or group of test-takers) in the same group would not get the same topic as the previous test-taker. Khabbazbashi (2017) says that in this kind of situation, most people would believe that even though the speaking test topics are different, they are all about the same level of difficulty. Despite having different topic familiarity backgrounds, one test-taker and another would appear to have the same chance of passing the test. In reality, this is a concern because tests with very unfamiliar topics could show that there might be bias against different groups of test-takers who are less familiar with the given topics. Studies which look into the area of effects of topic familiarity of L2 learners towards oral performance concerning L2 assessment have been growing worldwide in the past two decades (e.g. Huang et al., 2018; Khabbazbashi, 2017; Lumley & O'Sullivan, 2005; Qiu, 2019). Unfortunately, relatively few studies have focused on how Malaysian test-takers' topic familiarity could influence their oral test scores. To date, studies on topic familiarity of Malaysian students have revolved around its connection to L2 reading anxiety (Rajab et al., 2012) and L2 listening anxiety (Tahsildar & Yusoff, 2014); studies focusing on L2 oral performance such as that conducted by Abu Kassim & Zubairi, (2006) and Mohd Noh & Mohd Matore (2022) are very scarce. Therefore, in order to address this gap, it is crucial to recognise the effects of topic familiarity on the speaking performance of Malaysian test-takers. This understanding is vital for enhancing the design of the test task and refining the approach to its administration.

The following are the research questions which guide the study:

1) To what extent are the test-takers' speaking test measures valid and reliable considering:
   a. Item fit statistics
   b. Item separation
   c. Category functioning
2) How do item difficulty measures differ according to level of topic familiarity?

Literature Review

In this section, an overview of topic familiarity is presented. Subsequently, past studies concerning topic familiarity in relation to oral performance assessment and previous research in the Malaysian context is discussed.

Topic Familiarity

Topic familiarity is a variable representing individual characteristics of the test-taker within the language testing contexts which may have been overlooked by many tests (Banerjee, 2019; Brown, 2003; Huang et al., 2016; Qiu, 2019). Over the course of the last few decades, topic familiarity in language testing has been investigated in previous studies (e.g., Abu Bakar et al., 2019; Afaf Ayed Alrowaithy, 2021; Fu et al., 2021; Ovilia, 2019; Xiaolei et al., 2023). Past findings indicate that possessing knowledge or experience in a particular field may have a positive influence on one's ability to perform well in assessment evaluating inter-language fluency, complexity, and accuracy. In a study conducted by Banerjee (2019), the researcher examined the impact of topic familiarity on the performance of second language (L2) learners in language assessment. The research also discovered that the language proficiency performance of second language learners is influenced by their understanding of content and vocabulary, despite the fact that these factors manifest distinctively in the processes of acquiring and sharing information. This finding aligns with the proposition put forth by Skehan (2014) that familiarity with a given topic might aid in the process of conceptualisation.

Topic Familiarity and Speaking Assessment Performance

Research has been given due attention to the validity of rubric-based performance assessments of speaking (e.g., Bijani & Khabiri, 2017; Hidri, 2018; Huang et al., 2018; Mohd Noh & Mohd Matore, 2022; O'Grady, 2019; Sims et al., 2020), with only a handful of studies exclusively focusing on the influence of topic familiarity on speaking assessment (Banerjee, 2019; Han & Md Yusof, 2019; Huang et al., 2018; Khabbazbashi, 2017; Qiu & Lo, 2016). As empirically evidenced by certain research studies, it was found that topic familiarity significantly affected oral performance among learners (e.g. Bui & Huang, 2016; Huang et al., 2018; Lumley & O'Sullivan, 2005; Qiu, 2019). Still, the influence of topic familiarity variably depends on other factors such as planning time, test-taking strategy, anxiety, and confidence (Abu Kassim & Zubairi, 2006; Ameri-Golestan, 2016; Bui & Huang, 2016).

In their study, Bui & Huang (2016) discovered that the level of familiarity with the subject of a language task had an impact on the results of language task planning for learners; task-internal readiness was proposed to help test-takers with fluency. The study conducted by Qiu (2019) provided evidence that a total of eight participants, accounting for 38.10% of the sample, reported perceiving benefits in the practise of reiterating novel topics.

Previous Studies on Topic Familiarity in Malaysia

Reviewing past studies in the local context, it was found that the topic familiarity facet has revealed to influence Malaysian tertiary level learners where the topics for speaking assessment are usually selected among familiar topics (Abu Bakar et al., 2019; Lai Kuen et al., 2017). Willingness to communicate among Malaysian students were reported to be influenced by factors including topic relating to one's experience and the fear of being ridiculed or judged (Jahedi & Ismail, 2020) . This could explain why Malaysian students in the classroom would not be able to keep up with certain parts of a conversation or discussion in the classroom due to lack of topic familiarity towards certain topics discussed by their teacher or lecturer (Hiew, 2012). Lack of comprehension towards content would definitely affect the learners when they sit for their language proficiency assessments; in turn, this will apparently be connected to what is taught during classroom lessons.

The Malaysian context has seen studies concerning effects of confidence and anxiety learner factors and task aspects of task difficulty of institutional speaking tests (e.g. Abu Kassim & Zubairi, 2006; Ahmad Tarmizi et al., 2022; Idris & Zakaria, 2016; Lateh et al., 2015), studies on effects of rater characteristics towards speaking assessment (e.g. Baharudin et al., 2022; Mohd Noh & Mohd Matore, 2020) as well as learner-driven speaking assessment (e.g. Idris & Abdul Raof, 2019; Lateh et al., 2015), yet it is necessary to investigate additional factors that may have a role in test-takers' level of speaking performance within the context of Malaysia. In view of the paucity of research in this area and the importance assigned to the linked aspects of topic familiarity, a detailed investigation of this test-taker characteristics' impact on speaking ability is necessary and serves as the justification for the current study.

Concerning the current MUET test format, varying levels of topic difficulty may affect test fairness for test-takers without topic knowledge (Huang et al., 2018; Ketworrachai & Sappapan, 2022). The Malaysian Examination Council specified that MUET's latest CEFR-aligned speaking test topics would range from “familiar topics within the test-takers’ personal experience to more abstract topics in a range of fields and interest areas that may be encountered in late secondary/early tertiary education contexts” and “in the case of the more abstract topics, the test is limited to covering familiar and unfamiliar topics in different academic areas that a non-specialist would be able to talk about” (Malaysian Examinations Council, 2019, p. 21). Thus, the newest MUET speaking test specifications say that the test includes “unfamiliar topics,” which necessitates studying factors that may cause task performance variability.

Therefore, present study seeks to further examine the relationship between topic familiarity and oral performance among form six pre-university test-takers in Malaysia. This is in response to the paucity of research on the subject in the Malaysian educational context to date. By doing so, the study aims to reinforce and verify the critical importance of topic familiarity in speaking assessment.

Method

The study was conducted in a quantitative descriptive research design implementing purposive sampling. Participants responded to two different versions of the MUET speaking test using a parallel forms reliability design (each set covered five topics). The familiarity of the participants with the topics was assessed using topic familiarity self-reporting questionnaires, and their oral performances were rated by seven different raters. The data that were obtained from this were analysed using MFRM.

Context of Assessment

The assessment context for this study is the MUET Speaking module, a high-stakes standardised test of English proficiency. It consists of two face-to-face tasks required to be taken in a group of four test-takers; the first task is an individual presentation that is required to be delivered to the examiners and fellow candidates while the second task is a group discussion. The group discussion makes MUET unique since international speaking tests of IELTS, ACTFL, and TOEFL do not have this task in their respective speaking test formats (Idris & Abdul Raof, 2019).

With an information-transfer focus, this speaking test is primarily made up of independent, heavily topic-based tasks. These tasks need test-takers to rely on their background knowledge (Khabbazbashi, 2017) in order to reply to prompts and generate topics. Due to the strict examiner frame, candidates cannot choose topics or manage them (Malaysian Examinations Council, 2019).

Thus, it is possible to argue that the influence of topic familiarity seems notable in this test, and therefore this setting was selected for the study.

Participants

The study involved two groups of participants, which were the test-takers and the raters. Test-takers in this study were 40 Malaysian speakers of English as a Second Language aged between 18 and 20. There were 26 females and 14 males. All were enrolled in Form Six which made it mandatory for them to take MUET preparatory classes in their respective schools. They therefore constituted a fairly homogeneous sample in terms of L1, cultural background and exposure to English as a second language.

For the second phase, the rating of the speaking assessment tasks by the test-takers was carried out by seven raters, consisting of five females and two males who spoke English as a second language. They were chosen for the position based on their academic credentials, as well as their vast experience in both teaching and assessing a range of speaking assessments that are CEFR-aligned. It is important to note that all of the raters were provided training prior to evaluating the speaking tasks.

Research Instruments

Speaking tasks were selected from a collection of MUET previous papers that were made available to the general public. In order for (any) differences in scores to be predominantly attributable to topic differences and test-takers' topic familiarity, it was crucial to ensure that, with the exception of task topic, all other task-related variables were controlled for (Bachman, 2002; Weir et al., 2006) type of task input was controlled for by following the MUET speaking test format, the examiner role was fulfilled by a trained MUET teacher to control for any interlocutor effects (O’Sullivan, 2000), and by complying attentively to the MUET administration guidelines, the researcher was able to ensure the quality of the test's delivery, except for the condition of having one examiner instead of two in the real MUET speaking test.

To make sure that all the topics included in the study were able to be estimated in the study, an incomplete-connected data collection design (Eckes, 2009; Weir & Wu, 2006) was adopted; two test versions were applied (Q and R) each consisting of 4 Task A topics and one Task B topic following the MUET Speaking Test format. Two common tasks, which were the Task B tasks, were used in versions Q and R in order to create the necessary *common link* between the tests, allowing for coverage of 10 different topics (see Table One), meeting the requirements of MFRM.

According to the MUET speaking test, participants take the role of Candidate A, Candidate B, Candidate C, and Candidate D. Each candidate responded to versions Q and R resulting in 4 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 (B), there was full overlap on all task types within each group which ensured construct coverage and supported quality of the equating.

Rating Scale

The rating scale in this study was utilised by theraters to score the speaking performance of the test-takers in the video recordings. The rating scale was adapted from a CEFR-aligned rating scale which was developed, validated, and currently being used in the 17 Malaysia matriculation college institutions all over Malaysia to score course based speaking assessments which is modelled from the MUET. The original 6-band rubric was modified to a 5-band one. This five-band analytic scale consists of four criteria: Task Fulfillment (TF), Language (L), Communicative Ability (CA), and Group Discussion (GD). Scores were awarded for each criterion (as whole bands) to be estimated in the FACETS software.

Questionnaire on Topic Familiarity

According to the viewpoint of this research, differences in test-takers' topic familiarity of a task are a complex characteristic that cannot be simply predicted. After the speaking tests, participants completed out topic familiarity questionnaires to gauge their level of topic familiarity. A four-point Likert scale was used to collect answers to the questionnaire's four questions.

Speaking Performance Scores

Using the FACETS software, the MFRM analysis of speaking test scores was conducted (Linacre, 2023). The researchers began by conducting a number of MFRM analyses runs with four facets, using test-takers, raters, tasks, and items as the facets. The resulting subject measurement reports were utilised in answering Research Question (i). After that, a MFRM run with five facets was analysed, and the topic familiarity was added. The topic familiarity measures were categorised into four categories: least familiar, less familiar, familiar, and most familiar. The Topic familiarity measurement summary was then used to answer RQ (ii).

Research Procedures

Participant data collection took place in the Form six classes during school hours. The participants were called out group by group to a private room as the mock exam hall, where two versions of the speaking test were administered in succession. Once the test-takers were finished with all four tasks (1 topic for each task), test-takers then completed the topic familiarity questionnaire to measure their topic familiarity based on the 4 topics that they had to prepare and perform for.

Responses to topic familiarity questionnaires were scored. Each recorded speaking test was edited and the sequence of videos was arranged providing 20 speaking files (10 groups × 2 versions) for the raters to score. The researcher shuffled the sequence of the videos to make sure each rater had different arrangements of videos. This was done to minimise potential rater halo effect. Rater training was provided via a teleconference platform with all the raters. Following Weir & Wu (2006) , a complete judging plan was used where all test-takers were judged by all raters.

Table 1. *Incomplete-connected data collection design*

Topics	Candidate A	Candidate B	Candidate C	Candidate D

	Task Types	Test-takers: 1, 5, 9, 13, 17, 21, 25, 29, 33, 37	Test-takers: 2, 6, 10, 14, 18, 22, 26, 30, 34, 38	Test-takers: 3, 7, 11, 15, 19, 23, 27, 31, 35, 39	Test-takers: 4, 8, 12, 16, 20, 24, 28, 32, 36, 40
A.1	1	×			
A.2	1		×		
A.3	1			×	
A.4	1				×
B.1	2	×	×	×	×
A.5	1	×			
A.6	1		×		
A.7	1			×	
A.8	1				×
B.2	2	×	×	×	×

Test Version Q topics: A.1, A.2, A.3, A.4, B.1

Test Version R topics: A.6, A.7, A.8, A.9, B.2

Data Analysis

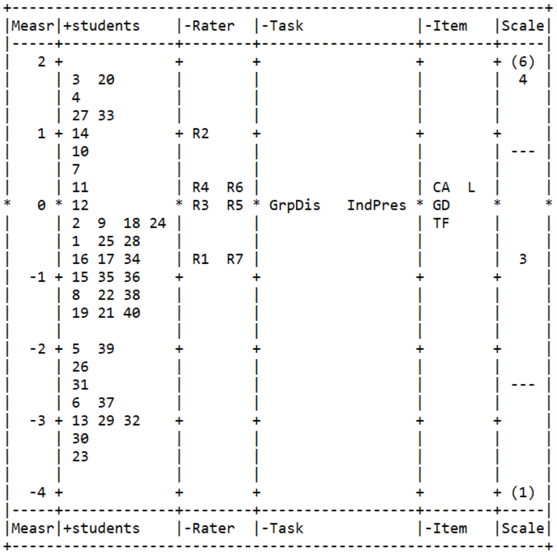
The MFRM analysis of speaking scores was carried out using FACETS(Bonk & Ockey, 2003) (Linacre, 2011). A series of four-facet MFRM analyses with examinees, raters, topics and criteria as facets were first run. Resultant topic measurement reports were used to address RQs (i) and (ii). A five-facet MFRM was subsequently run where topic familiarity was conceptualised as an additional facet.

Results

The wright map (Figure One) visually represents MFRM, displaying calibrations for all facets in the first FACETS software run. The test-taker measurement report illustrated a wide distribution of speaking abilities spanning about 5 levels from -3.58 – 1.73 logits. Separation indices indicated about eight statistically discriminating speaking ability strata (G=8.24, H=11.31), with a Rasch person separation reliability of r=.99. The seven raters in the study demonstrated high levels of consistency in their marking, with infit statistics lying between lower and upper control limits of 0.7-1.3., suitable for tests of lower stakes (Aryadoust et al., 2021). The results of the criterion facet suggested that the analytic criteria used in the study (TF, CA, L, and GD) contributed in various ways to classifying test-takers into different ability levels. An examination of the structures and categories of rating scales revealed that the categories within the scales were generally functional.

Test-takers' speaking test measures validity and reliability

For the first stage, a 3-Facet MFRM run was conducted in order to estimate the validity and reliability of the test-takers' speaking performance measures. Item fit statistics, item separation, scale functioning, and threshold gaps were first investigated to make sure the data is fit before the second MFRM run was employed.



```
+--------------------------------------------------------------+
|Measr|+students    |-Rater  |-Task             |-Item   |Scale|
|-----+-------------+--------+------------------+--------+-----|
|   2 +             +        +                  +        + (6) |
|     | 3  20       |        |                  |        |  4  |
|     | 4           |        |                  |        |     |
|     | 27 33       |        |                  |        |     |
|   1 + 14          + R2     +                  +        +     |
|     | 10          |        |                  |        | --- |
|     | 7           |        |                  |        |     |
|     | 11          | R4  R6 |                  | CA  L  |     |
*   0 * 12          * R3  R5 * GrpDis   IndPres * GD     *     *
|     | 2  9  18 24 |        |                  | TF     |     |
|     | 1  25 28    |        |                  |        |     |
|     | 16 17 34    | R1  R7 |                  |        |  3  |
|  -1 + 15 35 36    +        +                  +        +     |
|     | 8  22 38    |        |                  |        |     |
|     | 19 21 40    |        |                  |        |     |
|     |             |        |                  |        |     |
|  -2 + 5  39       +        +                  +        +     |
|     | 26          |        |                  |        |     |
|     | 31          |        |                  |        | --- |
|     | 6  37       |        |                  |        |     |
|  -3 + 13 29 32    +        +                  +        +     |
|     | 30          |        |                  |        |     |
|     | 23          |        |                  |        |     |
|     |             |        |                  |        |     |
|  -4 +             +        +                  +        + (1) |
|-----+-------------+--------+------------------+--------+-----|
|Measr|+students    |-Rater  |-Task             |-Item   |Scale|
+--------------------------------------------------------------+
```

Figure 1. Wright map (3-Facet MFRM) of test-taker, rater, and item (Criteria) difficulty

Item Fit Statistics

Item fit is one of the Rasch assumptions that needs to be fulfilled before further detailed analysis is performed. Fit indices are crucial in informing the study whether or not the assumption of unidimensionality of the measured construct is obtained (Bond & Fox, 2015). Report on item fit will ensure the items used in this study are able to measure speaking skills and traits among the test-takers. Item fit is determined by the value of infit and outfit for mean square (MnSq) and standardised fit statistics (Zstd).

Linacre (2002) outlines that the expected value for MnSq is 1.0, while Fisher (2007) states that value for MnSq between 0.77 to 1.30 logits are acceptable. In this study, the infit and outfit MnSq for the Communicative Ability criterion were 1.00 and 1.02 respectively; the values for the Language criterion (0.89 and 0.88 respectively), Group Discussion (1.06 and 1.06 respectively) and Task Fulfilment criterion (1.26 and 1.25 respectively) are indicated in Table Two. It is evident that the value for infit MnSq and outfit MnSq for all the four criteria used in this study are within the accepted range.

Next, the standard error indicates the accuracy of the measurement for each item. The standard error for all the four items are less than 0.25 and is adequate for its accuracy, as suggested by Fisher (2007). Specifically, the SE are ranged between 0.05 to 0.07. Table Two displays a summary of the fit statistics of all the items as criteria of the speaking test.

Table 2. *Report on item fit*

Item	Logit Value	S.E	Infit MnSq	Outfit Mnsq
Communicative Ability (CA)	0.19	0.05	1.00	1.02
Language (L)	0.19	0.05	0.89	0.88
Group Discussion (GD)	-0.05	0.07	1.06	1.06

Task Fulfillment (TF)	-0.33	0.05	1.26	1.25

Item Separation

Input on item separation is important to inform the study about the extent to which the items used are able to discriminate test-takers' capabilities. Item separation was analysed through separation ratio, index and reliability. Table 3 shows that the separation ratio for the four items used is 4.28. This signals that the difficulty of the items is divided into four levels of difficulty in relation to the precision of measure. Next, the value separation index is 6.04 which indicates that the items are able to discriminate the test-takers into six different levels of ability. Finally, the separation severity appears to achieve the desirable value, 0.95.

Table 3. *Item separation report*

Statistics	Item facet
Separation ratio	4.28
Separation index	6.04
Separation reliability	0.95

Scale Functioning

Analysis on scale functioning seeks to determine the quality of scoring scale categories used by raters. The objective of scale functioning is to identify whether the scales were able to measure the intended construct and also whether raters were able to use all the scales consistently. The findings emerged from the analysis can also inform the study if there is any problematic scale category which needs to be combined, divided or omitted (Myford & Wolfe, 2004). Linacre (2002) states that there are six basic assumptions about scale functioning that need to be fulfilled before further analysis is performed. Table Five illustrates the report of the six scales used in the study extracted from FACETS.

The first criterion conditions that each scale must be used more than ten times (Bond & Fox, 2005). Based on the findings shown in Table Five, it was found out that each scale in this study was used more than ten times. Specifically, scale 3 was the most frequently used scale by raters and awarded to test-takers 2,050 times (52%). This is followed by scale 2 that was used 935 times, scale 4 (734 times), scale 5 (143 times) and scale 1 (49 times). Only scale 6 did not meet the requirement with only 8 times used by the raters. This may be due to the reason that this high ability scale is only achievable by a minority of the test-takers in this context.

The second criterion is about the monotonicity of the average measures. It conditions that the average measure for each scale must be monotonically ascending. This criterion is fulfilled in this study as the findings reveal that the average measures start with -3.07 for scale 1, followed by -2.16 for scale 2, -0.81 for scale 3, 0.43 for scale 4, 1.31 for scale 5 and finally the average measure for scale 6 is 1.76. The increment pattern of the scales suggests that the scales were used with uniformity.

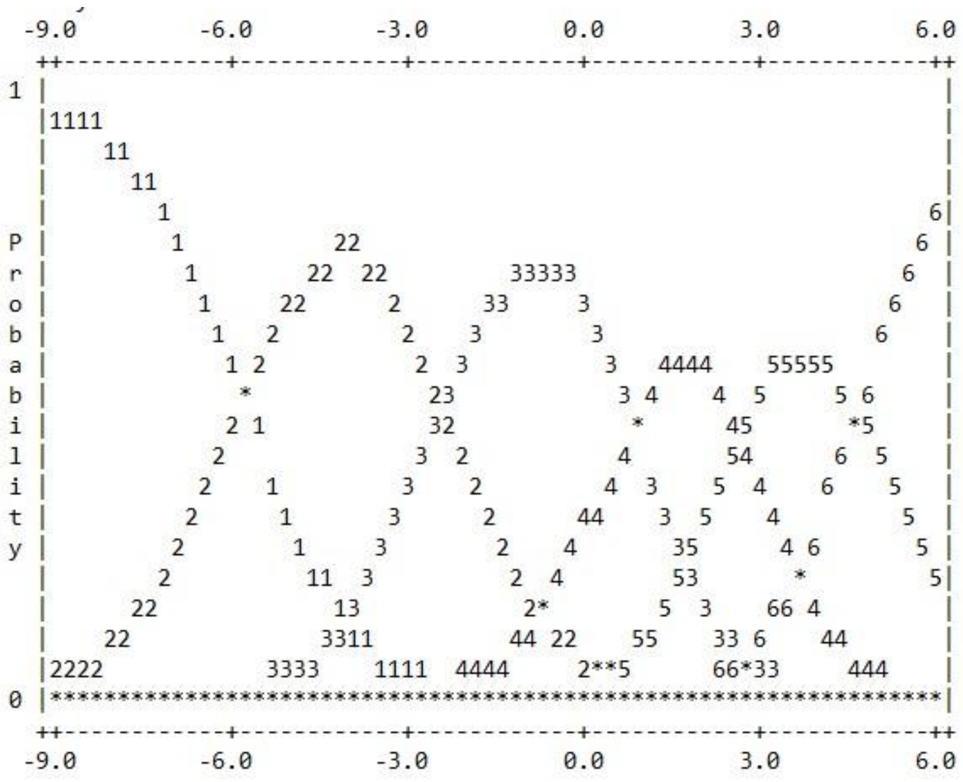
As for the third criterion, the infit mean square for each scale must be less than logit 2.0. The results in Table Five portrays the value of outfit MnSq for each scale are within the acceptable range that is either 1.0 or 1.1.

Next, the fourth criterion requires the threshold measure to be increasing from the smallest to the biggest scale. This assumption is also fulfilled as the thresholds for each scale in this study ascend systematically starting with -5.67, followed by -2.31, 0.88, 2.62 and finally 4.48. These results are important to prove the absence of disordered scales in this study.

Table 5. *Threshold gaps*

Scales	Gap	Threshold
$S_{1\text{-}2}$	0.00 – (-5.67)	5.67
$S_{2\text{-}3}$	-5.67 – (-2.31)	1.00 < 3.36< 5.00
$S_{3\text{-}4}$	-2.31 – 0.88	1.00 < 3.19< 5.00
$S_{4\text{-}5}$	0.88 – 2.62	1.00 < 1.74< 5.00
$S_{5\text{-}6}$	2.62 – 4.48	1.00 < 1.86< 5.00

Finally, the last criterion aims at determining the distinctions between thresholds through a visual representation of the scale usage given by the output. This was conducted through observation of the curve graph produced from the FACETS analysis. The assumption is fulfilled when each category has a distinct peak in the probability curve graph. When there is even one category with no clear peak, it signals that the particular scale is not sufficiently used by the raters. Figure Two portrays that all the scales in this study managed to have their own peak and there isn’t any scale that is hiding behind that of another. Hence, the sixth criterion is fulfilled in the study. To sum up, all the necessary six criteria were fulfilled in the study. Thus, all the six scales are maintained and further analysis was able to be executed for the second MFRM run positioning topic familiarity as a dummy facet.



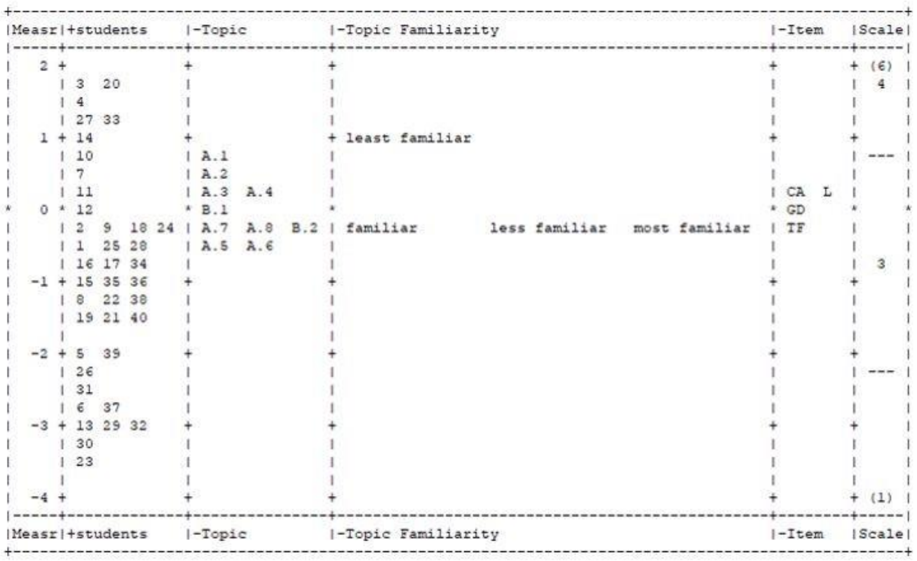
```
 -9.0         -6.0         -3.0          0.0          3.0          6.0
  ++------------+------------+------------+-------------+-----------++
1 |                                                                  |
  |1111                                                              |
  |    11                                                            |
  |      11                                                          |
  |        1                                                        6|
P |         1           22                                         6 |
r |          1        22  22         33333                        6  |
o |           1     22      2      33     3                      6   |
b |            1   2         2    3        3                    6    |
a |             1 2           2  3          3   4444    55555        |
b |              *             23            3 4    4  5     5 6     |
i |             2 1            32             *      45       *5     |
l |            2              3  2           4       54      6  5    |
i |           2    1         3    2         4  3    5  4    6    5   |
t |          2      1       3      2      44    3  5    4         5  |
y |         2        1     3        2    4       35      4 6       5 |
  |        2          11  3          2  4        53       *         5|
  |      22             13            2*        5  3    66 4         |
  |    22              3311          44 22    55    33 6    44       |
  |2222            3333    1111  4444     2**5      66*33     444    |
0 |******************************************************************|
  ++------------+------------+------------+-------------+-----------++
 -9.0         -6.0         -3.0          0.0          3.0          6.0
```

Figure 2. Category probability curve graph

The Difference in Item Difficulty Measures when Differentiated by Test-takers' Level of Topic Familiarity

The study intends to identify if the difficulty of the items used in the speaking tasks appeared to be different when answered by test-takers of different topic familiarity. Figure Three depicts MFRM, displaying calibrations for all facets in the second run, anchoring the values for rater, task, and topic facets.

Based on the questionnaire on topic familiarity responses, the test-takers showed different levels of topic familiarity and they were divided into four different groups. The test-takers were either least familiar, less familiar, familiar or most familiar with the topics of the speaking test that they were assigned to respond to. Based on these different groups of test-takers familiarity, analysis was performed to determine if the item difficulty was distinct when differentiated by different group of test-takers. The findings revealed that the items were of different level of difficulty to different group of test-takers. Interestingly, the difficulty of the items has a linear pattern to the test-takers' familiarity of topics. Specifically, the items appeared to be most difficult for least familiar test-takers with logit value of 0.97. This is followed by the less familiar test-taker group with -0.29 logits for item difficulty and the familiar test-takers group with -0.32 logits.



```
+---------------------------------------------------------------------------------------------------+
|Measr|+students    |-Topic         |-Topic Familiarity                              |-Item   |Scale|
|-----+-------------+---------------+------------------------------------------------+--------+-----|
|   2 +             +               +                                                +        + (6) |
|     | 3  20       |               |                                                |        |  4  |
|     | 4           |               |                                                |        |     |
|     | 27 33       |               |                                                |        |     |
|   1 + 14          +               + least familiar                                 +        +     |
|     | 10          | A.1           |                                                |        | --- |
|     | 7           | A.2           |                                                |        |     |
|     | 11          | A.3  A.4      |                                                | CA  L  |     |
*   0 * 12          * B.1           *                                                * GD     *     *
|     | 2  9  18 24 | A.7  A.8  B.2 | familiar        less familiar   most familiar  | TF     |     |
|     | 1  25 28    | A.5  A.6      |                                                |        |     |
|     | 16 17 34    |               |                                                |        |  3  |
|  -1 + 15 35 36    +               +                                                +        +     |
|     | 8  22 38    |               |                                                |        |     |
|     | 19 21 40    |               |                                                |        |     |
|     |             |               |                                                |        |     |
|  -2 + 5  39       +               +                                                +        +     |
|     | 26          |               |                                                |        | --- |
|     | 31          |               |                                                |        |     |
|     | 6  37       |               |                                                |        |     |
|  -3 + 13 29 32    +               +                                                +        +     |
|     | 30          |               |                                                |        |     |
|     | 23          |               |                                                |        |     |
|     |             |               |                                                |        |     |
|  -4 +             +               +                                                +        + (1) |
|-----+-------------+---------------+------------------------------------------------+--------+-----|
|Measr|+students    |-Topic         |-Topic Familiarity                              |-Item   |Scale|
+---------------------------------------------------------------------------------------------------+
```

Figure 3. Wright map (4-Facet MFRM) of test-taker, topic, topic familiarity, and item (Criteria) difficulty

Finally, the item difficulty was the least difficult or the easiest for test-takers that are most familiar with the topic with the logit value -0.36. The Chi-Square analysis depicted that the differences of item difficulty differentiated by the test-takers' level of topic familiarity were statistically significant with the Chi-square value, $\chi^2 = 47.5$, $df = 3$, $p < 0.01$. Thus, the null hypothesis that there was not any difference in item difficulty based on their topic familiarity was rejected.

Table 6. *The topic familiarity measurement report*

Topic Familiarity Level	Obs avg	Fair-M avg	Measure	Model S.E.	Infit MnSq	Outfit MnSq
Least Familiar	2.4	2.55	0.97	0.19	1.21	1.20
Less Familiar	2.6	2.95	-0.29	0.06	0.95	0.95
Familiar	3.1	2.96	-0.32	0.04	1.07	1.06
Most Familiar	3.3	2.97	-0.36	0.05	1.08	1.09
Mean (N=4)	2.8	2.86	0.00	0.08	1.08	1.08
SD	0.4	0.20	0.65	0.07	0.11	0.10

Model, Sample: RMSE .10 Adj (True) SD .64 Separation 6.20 Strata 8.60 Reliability .97
Model, Fixed (all same) chi-square: 47.5 *df*:3 significance (probability): .00

To summarise, the results indicate that the speaking measures were both reliable and valid, as evidenced by their item fit statistics, item separation, and category functioning. Additionally, when the speaking measures were calibrated with the topic familiarity dummy facet, significant differences were observed between the various topic familiarity group categories in terms of the speaking test criteria items. In general, it was found that the Communicative Ability and Language criteria were the most challenging for the test-takers.

Discussion

Question 1: To what extent are the test-takers' speaking test measures valid and reliable considering: a) item fit statistics, b) item separation, c) category functioning

It was found that the speaking test measures were valid and reliable as an outcome of examining the three stated indicators. When evaluating internal consistency, Rasch is considered a better measure as it has the ability to transform the measurement units to logit in ratio data type. This transformation results in a linear numerical representation, which can be extremely helpful in analysing the data and drawing meaningful insights. The study chose to employ the Rasch model to measure validity and reliability of the data where it is more valuable in terms of the information it gives as compared to other techniques, i.e. the Cronbach's Alpha from classical test theory. This is in line with the study by Anselmi et al. (2019) where they discovered that the Rasch model is a modern measure of internal consistency. Studies concerning speaking assessment measures like Khabbazbashi (2017), Mohd Noh & Mohd Matore (2022), and O'Grady (2019) also implemented the same methods for validity and reliability. They reported using similar methods in MFRM to validate their speaking measures before calibrating their respective dummy facets. For future research investigating rater-mediated assessment, item fit statistics, item separation, and category functioning could be considered.

Question 2: How do item difficulty measures differ according to level of topic familiarity?

When calibrated with the speaking test measures of the test-takers, the findings demonstrated that test-takers of higher performance were among those who indicated topics as most familiar towards the topics given; conversely, test-takers with lower speaking performance were among the group of test-takers which specified topics to be least familiar. Although the logit ranges between the most familiar, familiar and less familiar test-taker groups were not as large as that between least familiar, nevertheless the results showed a consistent pattern between the continuum of the topic familiarity groups and the speaking performance, serving evidence that topic familiarity has a direct affect towards speaking performance. These results are similar to that of Huang et al. (2018), Qiu (2019), and Xu & Qiu (2021) which also identified a statistically significant role for topic familiarity in L2 speaking performance. Huang et al. (2018) argue that language tests are unfair when the rating scale assesses topic development in conditions where the objective is not to test topical knowledge in the first place. Thus, high stakes language tests should have an eye on the mechanism of topic selection for their speaking tests. Specifically, the MUET proficiency test speaking component in Malaysia could consider constructive upgrading in selecting topics which would be familiar to both high and low proficiency test-takers.

Concerning the item criteria comprised in the speaking rating scale sourced by the raters, the results found that the item criteria most difficult for test-takers to perform in are the communicative ablity and language criteria. Having to sit for the speaking assessments in a group could have hindered the test-takers' fluency due to intimidation of being observed by fellow group members. In turn, heights of anxiety and a cutback in self-confidence would be most possible in a situation (Badrasawi et al., 2021) where the test-takers are obligated to speak about less familiar topics.

In addition, the language criteria would also be a challenge for lower profiency test-takers to perform well in if their cognitive capacity is incapable of processing both the conceptualisation of the least/less familiar topics and language accuracy simultaneously (Skehan, 2014). Pre-task planning time of 2 minutes before the Individual Presentation task could be too overwhelming for lower performers to prepare for a two-minute presentation; O'Grady (2019) suggests that 5 minutes could elicit substantial speaking performance. Bui & Huang (2016) suggested that task-internal readiness is applied to strategies of getting test-takers prepared for their language tests. In the context of the speaking test component, this study serves evidence that topic familiarity should be part and parcel of the preparation needed to enhance task-internal readiness that teachers can provide test-takers.

Conclusion

This study investigated how topic familiarity affects English Language proficiency speaking assessment, specifically of the MUET format, a high-stakes English test in Malaysia. Having advantageous levels of topic familiarity is widely acknowledged to facilitate in second language assessments. However, in language assessment settings, empirical evidence indicating the need to take into account test-taker topic familiarity in designing speaking tasks is still inadequate. The overall effects of topic familiarity conveyed to be significant towards MUET test-takers' performance in their speaking test practice, where there is involvement of speaking production from familiar and less familiar topics in language testing contexts. Therefore, steps should be taken to lessen (any) negatively influential topics in order to facilitate test-taker readiness and ensure that test-takers perform at their optimum level; it could be recommended to provide a choice structure in which test-takers are subject to choose their topic for speaking test. As a result, test-takers would be able to display their true language proficiency when cognitive load can be optimumly used towards speaking about a topic that is familiar to them.

About the Authors:

Nurul Iman Ahmad Bukhari is teaching English at Universiti Malaysia Kelantan. She is also currently a PhD TESL candidate at Universiti Putra Malaysia. Her research interests include ESL speaking performance and language assessment. http://orcid.org/0000-0001-9838-2243

Lilliati Ismail, Ph. D. is a Senior Lecturer at the Faculty of Educational Studies, Universiti Putra Malaysia. Her research interests include grammar instruction and task-based language teaching. https://orcid.org/0000-0002-7977-7327.

Noor Lide Abu Kassim, Ph.D. is a Professor at the Kuliyyah of Education, International Islamic University Malaysia. Her research interests include Psychometrics and Education Evaluation. http://orcid.org/0000-0003-2328-7406

Abu Bakar Razali, Ph.D. is an Associate Professor at the Faculty of Educational Studies, Universiti Putra Malaysia. His research interests include teaching of English as a second language (TESL), literacy and language education. http://orcid.org/0000-0002-3181-1004

Nooreen Noordin, Ph.D. is an Associate Professor at the Faculty of Educational Studies, Universiti Putra Malaysia. Her research interests include content-based instruction, technology-enhanced language learning, learning styles, motivation, and best teaching practices in ESL. http://orcid.org/0000-0002-4970-2682

Muhamad Firdaus Mohd Noh is an English primary school teacher. He is currently a PhD candidate at Universiti Kebangsaan Malaysia. His research interests include psychometrics and language assessment. http://orcid.org/0000-0002-5429-6789

References

Abu Bakar, N. I., Noordin, N., & Razali, A. B. (2019). Improving Oral Communicative Competence in English Using Project-Based Learning Activities. English Language Teaching, 12(4), 73–84. https://doi.org/10.5539/elt.v12n4p73

Abu Kassim, N. L., & Zubairi, A. M. (2006). Interaction Between Test-taker Characteristics, Task Facets and L2 Oral Proficiency Test Performance. Educational Awakening: Journal of the Educational Sciences, 3(2), 139–159. http://irep.iium.edu.my/29551/1/29551_interaction_between_test_taker_characteristics.pdf

Afaf Ayed Alrowaithy. (2021). The Effect of Topic Familiarity on Improving EFL Saudi Female Students` Reading Comprehension. Journal of Educational and Psychological Sciences, 5(7), 118–135. https://doi.org/10.26389/ajsrp.b020620

Ahmad Tarmizi, S. A., et al. (2022). A Socio-Cognitive Perspective on the Factors Affecting Malaysian Business Students' Learning when Spoken in English in a Second-Language Classroom. International Journal of Learning, Teaching and Educational Research, 21(1), 67–91. https://doi.org/10.26803/ijlter.21.1.5

Ameri-Golestan, A. (2016). Effects of Teaching Strategies and Topic Familiarity on Persian Speaking IELTS Candidates' Speaking Scores. Modern Journal of Language Teaching Methods, 6(3), 44–50. https://www.webofscience.com/wos/woscc/full-record/WOS:000379367100004

Anselmi, P., Colledani, D., & Robusto, E. (2019). A Comparison of Classical and Modern Measures of Internal Consistency. Frontiers in Psychology, 10(December), 1–12. https://doi.org/10.3389/fpsyg.2019.02714

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. Language Testing, 38(1), 6–40. https://doi.org/10.1177/0265532220927487

Bachman, L. F. (1999). Appendix: Language testing – SLA research interfaces. In L. F. Bachman & A. D. Cohen (Eds.), Interfaces between Second Language Acquisition and Language Testing Research (pp. 177–195). Cambridge University Press. https://doi.org/10.1017/CBO9781139524711.010

Badrasawi, K. J. I., Kassim, N. L. A., Zubairi, A. M., Johar, E. M., & Sidik, S. S. (2021). English Language Speaking Anxiety, Self-Confidence and Perceived Ability among Science and Technology Undergraduate Students: A Rasch Analysis. Pertanika Journal of Social Sciences and Humanities, 29(3), 309–334. https://doi.org/10.47836/pjssh.29.s3.16

Baharudin, H., Maskor, Z. M., & Matore, M. E. E. M. (2022). The raters' differences in Arabic writing rubrics through the Many-Facet Rasch measurement model. Frontiers in Psychology, 13, 7760. https://doi.org/10.3389/FPSYG.2022.988272/BIBTEX

Banerjee, H. L. (2019). Investigating the Construct of Topical Knowledge in Second Language Assessment: A Scenario-Based Assessment Approach. Language Assessment Quarterly, 16(2), 133–160. https://doi.org/10.1080/15434303.2019.1628237

Bijani, H. (2018). Effectiveness of a Face-to-Face Training Program on Oral Performance Assessment : The Analysis of Tasks Using the Multifaceted Rasch Analysis. 5(4), 27–53. https://doi.org/10.30479/jmrels.2019.10667.1335

Bijani, H. (2019). Evaluating the effectiveness of the training program on direct and semi-direct oral proficiency assessment: A case of multifaceted Rasch analysis. Cogent Education, 6(1), 1670592. https://doi.org/10.1080/2331186X.2019.1670592

Bijani, H., & Khabiri, M. (2017). Investigating the effect of training on raters' bias toward test takers in oral proficiency assessment: A FACETS analysis. Journal of Asia TEFL, 14(4), 687–702. https://doi.org/10.18823/asiatefl.2017.14.4.7.687

Bond, T. G., & Fox, C. (2015). Applying the Rasch Model; Fundamental Measurement in the Human Sciences | Request PDF (3rd ed.). Routledge. https://www.researchgate.net/publication/312296223_Applying_the_Rasch_Model_Fundamental_Measurement_in_the_Human_Sciences

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. Language Testing, 20(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. Language Testing, 20(1), 1–25. https://doi.org/10.1191/0265532203lt242oa

Bui, G., & Huang, Z. (2016). L2 fluency as influenced by content familiarity and planning : Performance , measurement , and pedagogy. Language Teaching Research, 22(1), 94–114. https://doi.org/https://doi.org/10.1177/1362168816656650

Chonghui, L. (2019). Boosting English Standards. The Star Online. Available at https://www.thestar.com.my/news/education/2019/06/09/boosting-english-standards

Council of Europe. (2001). the Common European Framework of Reference for Languages : Learning, Teaching, Assessment. Council of Europe. https://doi.org/10.1017/S0267190514000221

Fischer, J. (2020). The underlying action-oriented and task-based approach of the CEFR and its implementation in language testing and assessment at university. Language Learning in Higher Education, 10(2), 301–316. https://doi.org/10.1515/cercles-2020-2021

Fu, J. S., Yang, S. H., & Yeh, H. C. (2021). Exploring the impacts of digital storytelling on English as a foreign language learners' speaking competence. Journal of Research on Technology in Education, 0(0), 1–16. https://doi.org/10.1080/15391523.2021.1911008

Geranpayeh, A., & Ahmad Zufrie, A. R. (2018). The Alignment Of Malaysian University English Test To The CEFR. 6th British Council New Directions in English Language Assessment Conference: Standards in Learning Systems, October, 21. Retreived from https://www.britishcouncil.my/sites/default/files/nd_booklet_16_oct_v1.pdf

Han, L., & Md Yusof, M. A. (2019). Students' Self-reflections of Own Participation in English Language Oral Class. LSP International Journal, 6(2), 79–92. https://doi.org/10.11113/lspi.v6n2.92

Hidri, S. (2018). Assessing spoken language ability: A many-facet Rasch analysis. Second Language Learning and Teaching, 9783319628837, 23–48. https://doi.org/10.1007/978-3-319-62884-4_2

Huang, H. T. D., Hung, S. T. A., & Hong, H. T. V. (2016). Test-Taker Characteristics and Integrated Speaking Test Performance: A Path-Analytic Study. Language Assessment Quarterly, 13(4), 283–301. https://doi.org/10.1080/15434303.2016.1236111

Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. Language Testing, 35(1), 27–49. https://doi.org/10.1177/0265532216677106

Idris, M. B., & Abdul Raof, A. H. Bin. (2019). Learner-Driven Oral Assessment Criteria for English Presentation. Journal of Nusantara Studies (JONUS), 4(1), 365. https://doi.org/10.24200/jonus.vol4iss1pp365-383

Idris, M., & Zakaria, M. H. (2016). Gauging esl learners' cefr ratings on oral proficiency in rater training. Man in India, 96(6), 1675–1682. Available at https://scholar.googleusercontent.com/scholar?q=cache:Rw6rZgYUSYUJ:scholar.google.com/&hl=id&as_sdt=0,5&scioq=GAUGING+ESL+LEARNERS'+CEFR+RATINGS+ON+ORAL+PROFICIENCY+IN+RATER+TRAINING+Mardiana+Idris+and+Mohamad+Hassan+Zakaria

Ismail, L., & Abd. Samad, A. (2014). Using Tasks and Repair Practices in the Malaysian L2 Classroom. The Asia-Pacific Education Researcher, 23(3), 499–510. https://doi.org/10.1007/s40299-013-0124-7

Jahedi, M., & Ismail, L. (2020). Factors affecting ESL students' willingness to communicate in english classroom discussions and their use of linguistic strategies. Universal Journal of Educational Research, 8(8), 3360–3370. https://doi.org/10.13189/ujer.2020.080808

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. Language Testing, 34(1), 23–48. https://doi.org/10.1177/0265532215595666

Lai Kuen, G., Rafik-Galea, S., & Swee Heng, C. (2017). Effect of Oral Communication Strategies Training on the Development of Malaysian English as a Second Language Learners' Strategic Competence. International Journal of Education and Literacy Studies, 5(4), 57. https://doi.org/10.7575/aiac.ijels.v.5n.4p.57

Lateh, N. H. M., Shamsudin, S., & Said, S. M. (2015). Influence of Malaysian University English Test Training on the Speaking Performance of Pre-University Students. Advanced Science Letters, 21(7), 2463–2465. https://doi.org/10.1166/asl.2015.6311

Linacre, J. M. (2023). Facets computer program for many facet Rasch measurement (Version 3.87.0). https://www.winsteps.com

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. https://doi.org/10.1191/0265532205lt303oa

Malaysian Examinations Council. (2019). Malaysian University English Test (MUET). In Batu Caves: Percetakan Warni. Retreived from https://www.mpm.edu.my/images/dokumen/calon-peperiksaan/muet/regulation/Test_Specification_Regulation.pdf

Mohd Noh, M. F. Bin, & Mohd Matore, M. E. E. Bin. (2020). Rating performance among raters of different experience through multi-facet rasch measurement (MFRM) model. Journal of Measurement and Evaluation in Education and Psychology, 11(2), 147–162. https://doi.org/10.21031/epod.662964

Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. Frontiers in Psychology, 13, 1-13 https://doi.org/10.3389/fpsyg.2022.941084

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. Language Testing, 36(4), 505–526. https://doi.org/10.1177/0265532219826604

Ovilia, R. (2019). The Relationship of Topic Familiarity and Listening Comprehension. 276(Icoelt 2018), 182–186. https://doi.org/10.2991/icoelt-18.2019.26

Qiu, X. (2019). Functions of oral monologic tasks: Effects of topic familiarity on L2 speaking performance. Language Teaching Research. 24(6), 745-764. https://doi.org/10.1177/1362168819829021

Qiu, X., & Lo, Y. Y. (2016). Content familiarity, task repetition and Chinese EFL learners' engagement in second language use. Language Teaching Research, 21(6), 681–698. https://doi.org/10.1177/1362168816684368

Rethinasamy, S., & Chuah, K. M. (2011). The Malaysian university English test (MUET) and its use for placement purposes: A predictive validity study. Electronic Journal of Foreign Language Teaching, 8(2), 234–245. https://doi.org/10.2139/ssrn.2146007

Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric Rating with MFRM versus Randomly Distributed Comparative Judgment: A Comparison of Two Approaches to Second-Language Writing Assessment. Educational Measurement: Issues and Practice, 39(4), 30–40. https://doi.org/10.1111/EMIP.12329

Skehan, P. (2014). Processing Perspectives on Task Performance. John Benjamins Publishing Company.

Tahsildar, M. N., & Yusoff, Z. S. (2014). Investigating L2 students' listening anxiety: A survey at a Malaysian university. International Journal of Language Education and Applied Linguistics, 1, 43–52. https://doi.org/https://doi.org/10.15282/ijleal.v1.418

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. https://doi.org/10.1191/0265532206lt326oa

Xiaolei, S., Ismail, L., Yurong, H., & Mengqi, W. (2023). Strategies of Content Knowledge Representation and EFL Learners' English Writing Proficiency: Mediating Role of Critical Thinking Skills. Arab World English Journal, 14(3), 309–323. https://doi.org/https://dx.doi.org/10.24093/awej/vol14no3.19

Xu, J., & Qiu, X. (2021). Engaging L2 Learners in Information-gap Tasks: How Task Type and Topic Familiarity Affect Learner Engagement. RELC Journal, 0(0). https://doi.org/10.1177/00336882211061628