International Journal of Instruction
e-ISSN: 1308-1470 ● www.e-iji.net

April 2020 ● Vol.13, No.2
p-ISSN: 1694-609X
pp. 677-690

Received: 03/12/2018
Revision: 02/12/2019
Accepted: 08/12/2019
OnlineFirst:01/02/2020

Developing an Alternative Listening Comprehension Test to Benchmark Malaysian Undergraduates' Listening Performances

Irma Ahmad
Academy of Language Studies, Universiti Teknologi MARA, Negeri Sembilan, Malaysia, *irma@ns.uitm.edu.my*

Mohamad Jafre Zainol Abidin
School of Educational Studies, Universiti Sains Malaysia, Malaysia, *jafre@usm.my*

Listening is a fundamental skill in which students are required to gain adequate proficiency for their successful academic achievement. Since 2010, reports from a national exam have exhibiting a worrying trend in listening component where most of the candidates scored level 1 and 2 which indicating them as limited and very limited users. In view of this situation, this study aimed to develop an alternative listening comprehension test to benchmark and profile learners' listening performance. Test constructed was based on Weir's socio-cognitive framework for validating tests and Geranpayeh and Taylor's Cognitive Processing Model. It consists of 50 items with 7 types of response formats and administered to 380 students from a public university. Students' performances were benchmark based on their test score and profiled according to 5 performance standards. From the findings, 37.4% of the participants cleared the listening performance benchmark where 142 met the expectations and 129 exceeded the expectations. A total of 102 of the participants fell below the expectation level. Benchmarking and profiling the students have offered comprehensive information on the students as early as possible in identifying who might be at risk or need help.

Keywords: benchmarking, listening comprehension test, listening performance, test, listening

INTRODUCTION

There is no doubt that listening skill is important in daily life as human beings spend substantial amount of their time involve in communication and other activities that required listening with it. Previous researchers had proven that people spend a lot of time for listening in their communication and academic purpose. In fact, a study by Imhof (2008) has confirmed the earlier studies that listening is required in about two-thirds of instructional time.

Citation: Ahmad, I., & Abidin, M. J. Z. (2020). Developing an Alternative Listening Comprehension Test to Benchmark Malaysian Undergraduates' Listening Performances. *International Journal of Instruction*, 13(2), 677-690. https://doi.org/10.29333/iji.2020.13246a

Despite the equal importance attached to both receptive skills (listening and reading) and productive skills (speaking and writing), listening is traditionally considered to be unimportant and was taken for granted. The assessment of listening is the least understood, least developed and yet, one of the most important areas of language testing and assessment (Khoii & Paydarnia, 2011). Many researchers discovered that listening has been taken for granted in the classrooms and students were not exposed to listening comprehension process as more attentions were given on other skills like reading and writing (Vandergrift, 2007; Field, 2008; and Selamat & Sidhu, 2011).

Even during entry test, listening ability is rarely taken into account and if it does, it is usually accorded as a minor importance and the score rarely describe the actual performance of the students. Field (2008) discovered large number of learners whose skills have been graded differently such as "intermediate" on grammar test yet "minimal" in listening skills during his visit to language schools as a listening researcher and an inspector. These learners blamed themselves for not being able to comprehend information in the classroom. When there is pressure on contact hours, it is often the listening session that is cut so that attention could be given more on other skills like reading and writing which lead to rarely assessed on listening skills and listeners pass undiagnosed (Field, 2008; Selamat & Sidhu, 2011; Robinson et.al, 2014).

BACKGROUND OF STUDY

In Malaysia, English language is learned as a second language and it is also the medium of instruction in most universities. Like many languages learning settings worldwide, listening is widely acknowledged as a neglected skill due to insufficient pedagogical development. It appears that other skills like reading and writing instruction and assessment have always been a prominent component in the Malaysian school curriculum. Findings from collaborative baseline project in 2013 by Ministry of Education (MOE) and Cambridge English Language Assessment (CELA) found that listening is neglected and given a little attention on both listening comprehension process and assessment at school. It was mostly due to insufficient opportunity to practice in and out of the classroom and the strong emphasis on reading and writing over listening and speaking found in the reviewed national curricula, assessment and learning materials (Robinson, Galaczi, Docherty, King, & Khalifa, 2014). In fact, another study revealed that listening has not been given the treatment or status in most English language learning classroom in Malaysia (Suchitra, Koo, & Kesumawati, 2014).

Currently, the Malaysian University English Test (MUET) classroom is the one and perhaps the only channel that provides treatment for listening in the language context. Even here, listening is tested as a separate skill and takes the form of practice tests, (Nair & Mathai, 2010). Focusing on model practice test does not help the students to improve their abilities to listen effectively in lecture halls, listening tests or in any academic setting that they may encounter in their tertiary education. Furthermore, reports from The Malaysian Examination Council indicate very poor performance on listening component of Malaysian University English Test (MUET) from the year 2010 until 2015 (Official Portal Malaysian Examination Council, 2010-2015). Majority of the candidates scored band 2 and below which indicated as limited users and very limited

users respectively. Another issue regarding testing listening is that there is no proper listening test descriptor available currently in the country that properly describes students' abilities. Most of listening test descriptor available describes the aggregated scores in general. For example, in MUET, with aggregate score of 100-139, a student is at band 2 and known as limited user and described as "not fluent, inappropriate use of language, very frequent grammatical errors" in his or her communicative ability.

The Ministry of Education (MOE) has set aspirational targets for 2025 by using the Common European Framework of Reference (CEFR) global scale in order to make it clear what can be achieved in each stage of our English language program. The target is now a national agenda and for undergraduate students the CEFR level is set at B2-C1. As MOE is using CEFR as a basis for reinforcing English language education in Malaysia, there is a need for a common international framework of reference for interpreting students' performance across universities. Since listening is one of the skills that has been taken for granted throughout school level, and mastering the listening skills is important for undergraduates to perform in both their academic and social life. Therefore, there is a need to develop an alternative listening comprehension test that is compatible and could be used to benchmark and profile undergraduates' listening ability so that their performance could be monitored and enhanced to fit the expectation of being undergraduate students. If their performance does not fit with the targeted level, any issues discovered while assessing the students' performance could be explored and further discussed so that possible solutions could be suggested in order to achieve the targeted level. Besides, once the students know how to listen to learn, they will not be left behind as they are able to comprehend lectures and participate in discussions.

OBJECTIVE OF THE STUDY

The aim of this study was to develop an alternative listening comprehension test that can be the main instrument to benchmark students' listening abilities. In view of the present situation which were inappropriate listening test and listening performance descriptors and listening assessment is being paid little attention, the first objective of this study is to construct a set of listening test instruments that are valid and reliable to be used in this study.

The second objective is to benchmark the students based on their test score and map their score to the listening performance descriptors so that their performance could be described based on the "can-do" abilities.

The third objective is to profile the students based on their listening abilities: whether the students are Primary Standard Performers, Secondary Standard Performers, Exceed Standard Performers, Comprehensive Standard Performers or Mastery Standard Performers.

METHOD

For the purpose of this study, quantitative research was applied as it allows more precise analysis and prediction. The choice of research design selected for this study was standardized testing and cross-sectional styles for collecting data. As a quantitative

research, it relies on operationalization of empirical data which is collected from a sample and translated into numerical form that can then be subjected to one or more statistical and claims are made about the "true" nature of the phenomenon under study (Bodie & Fitch-Hause, 2010). One of the limitations of standardized testing is that it cannot generalize reliability and validity across various age and cultural groups.

However, this study focuses on measuring outcomes, not on developmental change or trend for a period of time. Moreover, the main objective of this study is not intended to investigate participants' listening performances trend but to investigate the undergraduates' current performance at one particular time. The rationale for choosing this style is that it snapshots different samples at one or more points in time and enables different groups to be compared. Other reasons for selecting cross-sectional studies are that they are less expensive, relatively quickly to conduct and produce finding more quickly. Other than that, they are less likely to suffer from control effects and more likely to secure the co-operation of respondents on a "one-off" basis. Besides that, cross-sectional designs are able to include more subjects than cohort designs and large samples enable inferential statistics to be used to compare subgroups within the sample. Data from this study were collected at one point in time and then comparisons between programs of study were made.

Preliminary Study

In the beginning of this study, a preliminary study was conducted to 39 undergraduate students from various program of study with aim to look at factors affecting their listening performance and information on what type of listening test that will meet their requirement as a student. The preliminary study was conducted by distributing a questionnaire which was adapted from Gao (2014) entitled 'University Learners' Awareness of Listening Difficulties and Causes of Study'. From the findings, almost half of the respondents mentioned vocabulary and grammar are the problems that affecting their listening performance and followed by pronunciation, memory and background knowledge. Other than that, they would like to be exposed to a new way of testing listening with various speakers, situations and response formats. At the same time, they would like to have a test which will not affecting their overall test score, in other words they would like to have an informal test but still reliable to test their ability. They also agreed that by having a list of what they can do based on their score will help them to monitor their performance. Thus, based on preliminary study, it can be concluded that students' need for listening test has emerged from not just to have a good score and grade but also for their preparation to the actual world.

Population and Sampling Procedure

The overall population size for this study is 1170 students; nearly 1200 undergraduate students. By referring to sample size as suggested by Krejcie and Morgan (1970; as cited in Cohen, Manion and Morrison, 2005), the sample size recommended for population size of 1200 with sampling error of 5% and confidence level of 95%; is 285. Thus, by considering Krejcie and Morgan's recommendation, the target population for this study was 285 undergraduate students from five-degree programs at a public

university in Negeri Sembilan. However, the actual population was 380 undergraduate students as there were students who volunteered to participate in this study. The programs selected were offered at UiTM Negeri Sembilan and its branch campuses. The programs are Mass Communication and Media Studies (MC, n=239), Information Management (IM, n=80), Applied Sciences (AS, n=271), Sports Management (SR, n=210), and Corporate Administration (AM, n=370). The sampling strategy that was used in this study was probability sample as it can draw randomly from the wider population and has less risk of bias that a non-probability sample. The systematic sample was used in determining the sample size as it can have precision equivalent to random sampling (Fowler, 2009 as cited in Creswell, 2014).

Stages of Development

After considering the result of preliminary study, a research design which consists of eight stages with main and sub-activities was developed to collect data in this study. All the data collected at one point in time and comparisons were made and differences were highlighted. The stages cover from developing all the test instruments including Prescribed Listening Performance Descriptors (PLPD), Standardized Listening Comprehension Test (SLCT), cut scores and performance band chart. All the instruments were sent for expert judgement consents.

Development of prescribed listening performance descriptor (PLPD)

The main aim at this stage was to develop a standardized descriptor that was used to describe the respondent of this study. This descriptor described the respondents' listening ability by telling what they are able to do. There were five performances that were described by using "can-do" descriptions which were adapted from CEFR Overall Listening Comprehension Scale, DIALANG Listening Scale, EQUALS, TOEIC Can-Do Level Table and The ALTE Framework. Other elements denoted in this descriptor were listening micro skills from MUET and Rost's General Language Ability and Listening Ability.

Development of standardized listening comprehension test (SLCT)

Considering the findings from preliminary study on having various response format, speakers and situations, The Standardized Listening Comprehension Test (SLCT) was developed. In order to make sure that the standardized test has its psychometric properties to indicate the magnitude of its reliability and validity, a framework for developing and validating tests of listening by Weir (2005) were adapted and used in this study. As suggested by Weir, there are five key elements of validation framework that test developers need to address to ensure fairness. The key elements are context validity, theory-based validity, scoring validity, consequential validity and criterion-related validity.

A checklist was used to collect evidence on context validity of task setting and task demand. Theory-based validity was covered at literature review where the main theory grounded for this study were based on Buck's default listening construct (2001), Geranpayeh and Taylor's cognitive processing (2013), and Rost's general language

ability and listening ability (2011). For the purpose of scoring validity, item analysis-item discrimination and facility, internal consistency and grading were calculated.

The test consisted of four sections (A-D). In each of the sections, different skills were tested and adapted from MUET Listening Specification and Skills. In every section, students were required to respond to all items. The items tested 7 types of response formats which include sentence dictation, true and false, information transfer, short-answers, gap filling, matching responses and multiple-choice questions (MCQ). There were 5 speakers; 2 male and 3 females. The speakers are all Malaysian Malay and Chinese speakers and they mostly delivered the text by using standard spoken English language. A prototype listening comprehension test was developed and tested for the purpose of validity and reliability of the instrument. There were 50 questions and all the questions were given 1 mark for each correct answer.

For the purpose of validity and reliability of the instrument, a prototype of listening comprehension test was developed, pre-tested to 20 undergraduates and pilot testing it to 100 undergraduate students before run for actual test to 380 of the undergraduate students. All the comments given by the experts and students were used to improve the test and its instruments.

Development of cut-score for bands

Cut scores are selected points on the score scale of a test. The points are used to determine whether a particular test score is sufficient for some purpose. Cut-score for the Performance Band was established by comparing the listening performance of the high, average and low performers of the respondents from the piloting process. For the purpose of this study, the setting of cut score was reviewed by the same experts that viewed the SLCT and PLPD. In this study, experts were given sufficient information about what students can actually do rather than to depend on whether each judge happens to think what students can or cannot do. One of the information given to the experts were the CEFR Overall Descriptors of Listening Scale, the difficulties of the test questions as stated in the SLCT Test Specification and Listening Test Grid and Specification. Nevertheless, real data on the performance of actual students can help the experts to make more realistic judgements about the performance of borderline students.

The cut scores for Band 1 to Band 5 were developed based on the z-score. The respondents were categorized into the 5 bands which were generated based on their listening performance in the pilot test. After determining the cut scores of the performance bands, a "reality check" was made by the experts based on the pilot test scores so that more realistic judgements could be made. Table below shows the 5 bands and its scale of performance of listening performance chart and the score range based on the z-score.

Benchmarking and profiling listeners performance

For the purpose of this study to benchmark the students' performance, their scores were analysed and categorized into band 1 until band 5 as suggested in the Listening Performance Band. After the band is identified, participants' listening performance were

benchmarked by referring to Prescribed Listening Performance Descriptors (PLPD) and it uses the "can-do" descriptions in describing the performers' listening abilities.

After benchmarking the respondents' listening abilities, respondents' profile was decided whether they are at Primary Standard, Secondary Standards, Exceed Standard, Comprehensive Standard and Mastery Standard. Respondents who are in band 1 are profiled as Primary Standard Performer, band 2 as Secondary Standard Performer, band 3 as Exceed Standard Performer, band 4 as Comprehensive Standard Performer and band 5 as Mastery Standard Performer.

FINDINGS

The first objective of this study is to construct a reliable and valid standardized listening comprehension test. Buck (2001) suggested 0.60 for internal consistency and 0.90 for a higher stakes of listening test construction while Brown (2018) suggested 0.70 at a minimum for classroom testing, and .90 and greater for high-stakes testing. From the table below, the Cronbach alpha coefficient was found to be 0.867 for the actual study and 0.884 for the pilot study. Thus, this shows that 86.7% of the variability in the composite score is considered as true score variance or internally consistent reliable variance. Therefore, SLCT was a reliable test, consistent and dependable to be run for this study.

Table 1
Reliability Statistic for both Pilot and Actual Study

	Num. of students	Num. of Items	Cronbach's Alpha
Pilot Study	100	50	.884
Actual Study	380	50	.867

From the table 2 below, the mean score of the actual study is 32.49 marks with a standard deviation of 7.631. Based on the score, half of the students scored less than 33 marks. The minimum and maximum score are 12 and 46 respectively. While for the pilot study, the mean score is 33.71 with standard deviation of 8.277. Based on the score, half of the students scored less than 33 marks in the pilot test. The minimum and maximum score are 13 and 47 respectively. The range of marks for pilot and actual study was 34 and 33 respectively.

Table 2
Descriptive Statistic of Pilot and Actual Score

	N	Range	Min	Max	Mean	SD	Skewness		Kurtosis	
	Stat	Stat	Stat	Stat	Stat	Stat	Stat	SE	Stat	SE
Pilot Score	100	34	13	47	33.71	8.277	-.464	.241	-.313	.478
Actual Score	380	33	13	46	32.49	7.631	-.385	.125	-.467	.250

In order to indicate the normality distribution of actual and pilot test, the skewness and kurtosis, the Shapiro-Wilk Test, histograms and Normal Q-Q Tests were identified. Below is the descriptive table of Shapiro-Wilk Test that compare normality test for both actual and pilot study. Based on Shapiro-Wilk test, the distribution of the students' score in the actual and pilot study is not normal (SW=0.968), p-value=0.014< 0.05). Based on the skewness value of -0.458 in the actual study and -0.464 in the pilot study, the distributions are skewed to the left.

Table 3
Test of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Actual Study	.084	100	.081	.968	100	.014
Pilot Study	.084	100	.075	.968	100	.014

The skewness and kurtosis value were divided by their standard errors and produced z-value. The z-value should be in between -1.96 to +1.96. The skewness and kurtosis values for actual study is closely to 0 and not too large compare to their standard errors. The z-values for the skewness and kurtosis in actual study are -1.93 and -0.65 respectively. The values in both studies are in the span of -1.96 to +1.96. This indicates that the data scores in both studies are a little skewed and kurtotic but they do not differ significantly from normality. Therefore, in general the score in the actual were not normally distributed and a little bit skewed and kurtotic, but they do not differ too significantly from normality.

In determining the cut score for this study, the range of standard deviation used is 0.5 above and below the mean. Thus, 38% of the students are categorized as average performers with score-range between 29-36. The score-range that is in between 0.5-1.5 standard deviation above the mean score is 38-45 marks and 24% of the students are estimated to fall in this range. The score-range that is more than 1.5 standard deviation above the mean score is 46-50 and only 7% of the students was estimated to fall in this range. Table 5 shows the 5 band, the score range and its scale of performance.

Item analysis was conducted by calculating the value of item facility and item discrimination of test items. Overall item analysis of the test can be referred to Appendix 2. Table below shows the difficulty of test items in the actual study of SLCT. The table below was adapted from ScorePak® scoring process material that was taken from www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/

Table 4
Item Analysis of SLCT

Discrimination	Difficulty Hard (0-50)	Questions Number Medium (50-85)	Easy (85-100)
Poor (<0.1)	-	-	8, 45, 47, 48
Fair (<0.1-0.3)	6, 7, 9, 16, 22, 36	35, 40	13, 14, 24, 31, 46, 50
Good (>0.3)	2, 5, 15, 19, 20, 27, 28, 37, 38, 39	1, 3, 4, 18, 21, 25, 26, 29, 30, 32, 33, 34, 41, 42	10, 11, 12, 17, 23, 43, 44, 49

From the table, 8 items were considered as easy and good, 14 items as medium and good, 10 items as hard and good, 6 items as hard and fair, 2 items as medium and fair, 6 items as easy and fair, and 4 items as easy and poor.

For the second objective of the study, respondents' overall listening performance were benchmarked based on their performance from the test. Their performance is summarized in the table 5. From the findings, 37.4% of the respondents have met the listening performance expectations based on their score in the actual test. 109 or 28.7%

of the respondents were categorized as below expectations group with score below than 28 marks. 142 of them had met the expectations and 129 of the respondents had exceed the expectations. Table 5 below shows the overall respondents' performance in the test. Table 6 summarizes the respondents' profile as according to their program of study.

Table 5
Students' Overall Performances

Bands	1	2	3	4	5
Score Range	0-21	22-28	29-36	37-43	44-50
Scale Performance	Below Expectations		Meet Expectations	Exceed Expectations	
Total	37	72	142	104	25
N (380)	109		142	129	

For the final objective of this study, respondents are profiled based on their performance into five categories. From the findings, the lowest number of the respondents who fall into band 1 or are profiled as primary standard performers are from Corporate Administrative (AM) and followed by Mass Communication and Media Studies (MC) and Applied Sciences. However, there were 3 from Information Management (IM), 12 from Sport Management (SR) and 2 from Applied Sciences (AS) were profiled as primary standard performer. From 72 secondary standard performers, 20 of them were from SR followed by 19 from AS and 15 from AM Among the 142 exceed standard performers, 38 were from AS, 25 were from MC, 35 from AM, 26 from SR, and 18 from IM. Finally, for mastery standard performer, the highest number of respondents were 13 from AM, followed by 7 from MC, 2 from AS, 2 from SR and 1 from IM.

Table 7
Students' Profile

Program of Study	AM	IM	MC	SR	AS
Primary Standard Performers	2	10	1	19	5
Secondary Standard Performers	15	8	10	20	19
Exceed Standard Performers	35	18	25	26	38
Comprehensive Standard Performers	27	14	28	17	18
Mastery Standard Performers	13	1	7	2	2
N=380	92	51	71	84	82

Students who were profiled as a *Primary Standard Performer* generally have basic grammatical accuracy, vocabulary of words and phrases, and a very basic range of expressions about personal details which are enough and yet still require some effort to understand a simple information, conversations, instructions, directions and speech even in local accents. A *Secondary Standard Performer* has sufficient grammatical accuracy, vocabulary for expression and coping with survival needs which enable to deal with everyday situations with predictable content though sometimes misunderstanding occur and require to compromise the message especially in non-routine or unfamiliar situations and accents. As an *Exceed Standard Performer*, students usually have a reasonable grammatical accuracy, sufficient vocabulary to express most topics on everyday life and situations with predictable content though with some hesitation and limitation that require repetition when listening to extended speech, complex

instructions, unfamiliar topics and situations, and accents. While a *Comprehensive Standard Performer* has a good control and vocabulary range to be able to understand unpredictable situations, extended speech, lectures, complex instructions or directions within or outside the field, without too much effort as long as they are delivered in standard spoken language. Finally, a *Mastery Standard Performer* has a broad range of grammatical accuracy and vocabulary to be able to understand without too much effort the extended speech, unfamiliar topics or situations that are not clearly structured, or not in standard dialect or accents, with considerable degree of slang and idiomatic usage.

DISCUSSION

From the findings, the respondents' listening difficulties are varied at different levels of cognitive processing models. Respondents who are below expectation group mostly are having difficulty at the *input decoding, lexical search* and *syntactic parsing* of the Cognitive Processing Theory. They tend to make mistakes in isolating phonemic units from the basic sound waves, identifying words from the individual phoneme, and imposing a syntactic structure on group of words to produce utterances. These can be seen at respondents' answers at dictation (item number 1 until 5) and information transfer 2 (item number 36 until 43). For more complex tasks like map and direction and gap filling, these tasks involve the meaning construction process and discourse construction process. The task involves in contextualizing and enriching what they have heard with the real-world knowledge and inferences to create a full proposition representing what speaker really meant. Then, respondents have to take the new information on what they have just heard and incorporating it into a representation of the whole discourse linking to everything that has gone before. Respondents who scored 34 and above in the test or have met the expectation in band 3 and above, have actually achieved these two levels of cognitive processing. Therefore, the respondents' listening performance could be improved if they could recognize at which level of cognitive processing that they need to focus on. In fact, by using their score in this study, they could start planning on how to improve their listening performance in the future.

Although a majority of the respondents have met and exceeded the expectation level, attention should be given to respondents who were benchmarked at below expectation level. This is because, at tertiary level, almost all courses are taught in the English language and while listening to lectures or class discussion, students are expected to respond appropriately, to request the speakers to repeat what they said or to clarify what was said. Besides, based on a study conducted by Ho (2016), the improved listening proficiency among his participants have a statistically significant impact on their speaking, writing and reading proficiency. Other than that, learners developed efficient writing skills depending on their acquired knowledge of phonological, discourse organizations, syntactic structures and pragmatics of the language (Shanahan, 2006). Thus, it will be challenging to band 1 and band 2 students in comprehending the lectures and information based on their proficiency level. A band 1 student or a primary standard performer has basic listening skills which are enough but might still require some effort to understand a simple information, conversations, instructions, directions and speech even in local accents. Meanwhile, a band 2 student or a secondary standard performer

has sufficient listening skills to cope with survival needs which enable them to deal with everyday situations, with predictable content. However, sometimes misunderstandings occur which require them to compromise the message especially in non-routine or unfamiliar situations and accents.

As listening experiences are considered limited in band 1 and band 2, priority should be given to strategy training. Instructors could equip themselves with formulaic repair strategies. It is useful to demonstrate to them that they do not need to master everything what the speaker has said, understanding may fail but they should be shown that parts of input are redundant and can be ignored or avoided if not fully understood. This is the time where instructors should focus on building hypothesis around words that have been capitalized upon the process works which have been done on word recognition. A task-based approach can be recommended for secondary standard performers but instructors should consider of using a short and simple recording so that students can draw on general contextual cues but should not rely upon the assumption that they have built up a clear picture of the conversation as a whole. Instructors should consider of using comprehension tasks that are short, with a maximum of two voices instructional and conversational so that students could experience in dealing with different types of connected speech from the tasks given.

As listeners progress, achieving success at all levels of processing will become automatic. In other words, they do not require attention from the listeners. This is important for a listening test because it informs the processes which should be taught at a particular level. By referring to the Cognitive Processing model by Geranpayeh & Taylor (2013), low level listeners need to pay more attention to the lower-level processes, as they are not able to engage in higher level processing such as meaning construction or discourse construction. Therefore, by teaching higher processes to these candidates will not be appropriate. Similarly, in higher level learners, lower-level processes becoming automatic is expected. For these high-level learners, they need to be taught of high-level processes to suit their level of proficiency.

IMPLICATION OF THE STUDY

Assessment always influences the washback effect on goal of instruction and students' motivation (Rost, 2011). By assessing what it supposedly to assess, it provides positive feedback to the students on their development of listening ability and describe listening proficiency holistically. From the findings, majority of the respondents were profiled as exceed standards and above. However, there are still quite a number of them who were *Primary* and *Secondary Standard Performers* who are need to be given attention to. Thus, based on the students' profiles and their descriptions on what they can do, it could assist the lecturers, not just ESL lecturers, to improve their strategies in teaching listening skills or strategies in giving lectures including in selecting choices of vocabulary and phrases, rate of speech and even preparing their lecture notes to suit their students' abilities. They could slowly upgrade the vocabulary or use complex phrases or sentences until they meet the undergraduate students' level.

In terms of methodological aspect, this study has suggested a few response formats that were useful and yet forgotten by most educators in teaching English as second language

especially dictations. Dictation should be introduced again even at undergraduate level because students could use the dictation skill in their daily life activities such as writing important information from lectures or even from instructions given by their lecturers on important assignments or projects. Although in this study, dictation was tested by giving one sentence at a time and it does not seem to require the ability to understand inferred meaning; it clearly tested the respondents' short-term memory as well as writing ability. Besides, it also tests the respondents' understanding on how English sounds change in connected speech. With sufficient exposure to normal speed of English language acoustic input in their ESL context, it could help them in decoding and segmentation process of sounds into words. As highlighted in many literatures, these problems are primarily derived from insufficient knowledge and / or practice of complex English word variations such as re-syllabification (linking or liaison), reduction, assimilation, and / or elision which occur in normal-speed connected input (Field, 2008, 2004, 2003; Goh, 2000 and Kuo, 2010). Dictation may not important in testing, but it does give an impact to students' listening performance.

By using the findings from this study, it could be implied in teaching and learning listening comprehension in the classroom. Instructors could plan their lesson by paying more attention to the stages that are challenging or difficult for the students to master. Planning for lesson is crucial as instructors need to consider how to use the approaches to teaching listening in L2 classroom. Field (2008) suggested a multi-strand approach to L2 listening development, a lesson can be built based on the priorities of what students actually need to focus on. He suggested the combination of five strand approaches: processing training, strategy training, exposure to authentic speech, diagnostic activities and general comprehension work. For example, if the students are categorized as primary or secondary standard performers, the process of listening development that instructors should consider is lexical and decoding where students' ability to recognise words in connected speech by means of small-scale transcription exercises and tasks related to lexical segmentation, should be developed.

LIMITATION OF STUDY

Listening is incorporated in English language syllabuses from primary until post-secondary and undergraduate level of Malaysian education system. The study would be more comprehensive if it is studied at all levels of educational system. This is because it will give clearer picture on the issues of listening assessment if more comprehensive sampling could be conducted to all educational level and to both public and private universities in Malaysia. However, this study will be restricted to five groups of undergraduate students at one public university in Malaysia. This study only profiles the performance group of respondents at a given time. Other matters like listening instructions and strategies and preparing lectures materials to suit the needs of the students are not included in this study as these issues would entail to another study.

CONCLUSION

The listening skill plays a significance role in one's successful communication and academic achievement. It is a fundamental skill in which students need to gain adequate

proficiency so that they could participate in any form of communication. This study has achieved its aim to develop an alternative listening comprehension test to benchmark and profile undergraduates listening performance. In this study too, it has sought to respond to a number of issues related to how listening comprehension is planned, developed and administered. This includes the criteria that need to be considered in developing the test, how the results can be used to benchmark and profile the respondents and how different types of data or information of the study are brought to bear in improvising the process of teaching and learning listening. However, the findings and discussion are restricted to the five groups of undergraduate students in a public university and it only profiles the performance group of respondents at a given time. For future study, it is suggested that the number of participants be increased so that the findings can be generalized to all undergraduates in Malaysia. Other matters like listening instructions and strategies as well as preparing lectures materials to suit the needs of the students could also be included in future studies. Finally, from the findings and analysis, the implication of study was made to discuss on how this study could benefit the current way of testing listening.

REFERENCES

Bodie, G., & Fitch-Hausner, M. (2010). Quantitative research in listening: Explication and overview. In A.D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 46-93). Oxford: Blackwell.

Brown, G. T. (2018). *Assessment of student achievement.* New York: Taylor & Francis.

Buck, G. (2001). *Assessing listening* (Cambridge Language Assessment Series ed.). (J. Anderson, & L. Bachman, Eds.) Cambdrige: Cambdrige University Press.

Creswell, J. W. (2014). *Research design. qualitative, quantitative and mixed methods approaches.* California: SAGE Publications.

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal*, 325-334.

Field, J. (2004). An insight into listeners' problems: too much bottom-up or too much top-down? *System*, 363-377.

Field, J. (2008). *Listening in the language classroom.* Cambridge: Cambridge U.

Gao, L. (2014). *An exploration of L2 listening problems and their causes* (Unpublished doctoral dissertation). University of Nottingham.

Geranpayeh, A., & Taylor, L. (2013). *Examining listening: Research and practice.* Cambridge: Cambridge University Press.

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*(1), 55-75. http://dx.doi.org/10.1016/S0346-251X(99)00060-3.

Ho, S. H. (2016). The effects of listening comprehension on ESL learners' English language proficiency. *Malaysian Journal of ELT Research, 12*(2), 15-30.

Imhof, M. (2008). What have you listened to in School today? *International Journal of Listening, 22*(1), 1-12. doi:10.1080/10904010701802121.

Khoii, R., & Paydarnia, S. (2011). Test method facet and the construct validity of listening comprehension tests. *The Journal of Applied Linguistics, 4*(1), 99-121.

Lim, T. D. (2013). *Analyzing Malaysian English classroom: Reading writing, speaking and listening teaching strategies*. (C. o. Education, Producer, & University of Washington) Retrieved from http://digital.lib.washington.edu.

Lin, S. E. (2009). Malaysian students' English language reading standards: The case of Penang. Pulau Pinang: Universiti Sains Malaysia.

Majlis Peperiksaan Malaysia Laporan Tahunan 2013. (2013). Retrieved from Official Portal Malaysian Examination Council: portal.mpm.edu.my.

Malaysian Examination Council: MUET. (2015). Retrieved from Malaysian Examination Council: www.mpm.edu.my.

Nair, S., & Mathai, E. (2010). *A Survey of listening instruction in MUET classroom.* Universiti Teknologi MARA, Bureau of Research and Consultancy.

Official Portal Malaysian Examination Council. (2010-2015). Retrieved from STPM and MUET Examination Reports: portal.mpm.edu.my.

Robinson, M., Galaczi, E. D., Docherty, C., King, A., & Khalifa, H. (2014). Supporting national education reform: The Cambridge Malaysia Baseline Project. *Cambridge English: Research Notes*(58), pp. 40-44.

Rost, M. (2011). *Teaching and researching, listening.* UK: Pearson Education.

Selamat, S., & Sidhu, K. G. (2011). Student perceptions of metacognitive strategy use in lecture listening comprehension. *Language Education in Asia, 2*(2), 185-198.

Shanahan, T. (2006). Relations among oral language, reading and writing development. In S. G. C.A. Mac Arthur (Ed.), *Handbook of writing research* (pp. 171-183). New York, NY: The Guildford Press.

Suchitra, N., Koo, Y. L., & Kesumawati, A. (2014, March). Exploring the listening processes of pre-university ESL students. *Proc. Soc. and Beha. Sci, 118*, 475-472.

Understanding Item Analysis. (2018). Retrieved from Office of Educational Assessment http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis.

Vandergriff, L. (2007, February 22). *Listening: theory and practice in modern foreign language competence.* Retrieved from Subject Centre for Languages, Linguistics and Area Studies Guide to Good Practice: http://www.llas.ac.uk/resources/gpg/67#toc_1.