Malaysian Music Journal Vol. 5, Num. 1 (20-36)
ISSN 2232-1020

The Validation of a Basic Knowledge Test of Music for the Cultural Arts Guidance Program (PBSB) in Malaysia Using the 2 Parameter Logistic (2PL) Model Item Response Theory

Siti Eshah Mokshein[1], Zaharul Lailiddin Saidon[2] and Brian Doig[3]
[1]Faculty of Education and Human Development, Universiti Pendidikan Sultan Idris
[2]Faculty of Music and Performing Arts, Universiti Pendidikan Sultan Idris
[3]Deakin University, Australia
e-mail: eshah@fppm.upsi.edu.my[1], zaharul@fmsp.upsi.edu.my[2], b.doig@deakin.edu.au[3]

Abstract

This study is drawn from a larger study on the effectiveness of the Cultural Arts Guidance Program (PBSB) in Malaysian schools. It is a joint programme between the Department of National Culture & Arts (JKKN) and the Ministry of Education Malaysia. The PBSB effectiveness study was conducted in 2013 to help JKKN improve the programme implementation and set forth the future direction of PBSB. The three most popular areas of cultural arts, namely dance, music and theatre were studied. Several assessment instruments were developed based on the objectives of PBSB and the modules used in the programme. This study focuses only on the development and re-validation of the basic knowledge test of music used in the PBSB effectiveness study. The present article discusses the background of PBSB, some important findings from the PBSB effectiveness study and the psychometric characteristics of the items in the test from the perspective of the Item Response Theory (IRT). The test of multiple-choice items was administered to 437 PBSB students in primary and secondary schools that were selected through stratified random sampling technique. Data from the study were re-analysed using IRT to further establish the reliability and validity of the test. Overall, the test was found to possess sound psychometric characteristics as reflected by the model fit, the item-person map, reliability and validity of ability estimates and the difficulty, discrimination and guessing parameters. The test can be used to complement the existing assessment systems in PBSB, but different tests should be developed for each module.

Keywords music test, Item Response Theory (IRT), validation

INTRODUCTION

The Cultural Arts Guidance Program or PBSB is one of the core activities of the National Department of Art and Culture (JKKN), Malaysia. The PBSB was first introduced to several schools in 1996 as *Kumpulan Tunas Budaya* (KTB), focusing on dance. Following a Cabinet decision in March 2000 that the Ministry of Culture, Arts and Tourism should assist the Ministry of Education (MOE) to promote the Culture & Art Clubs in schools, KTB underwent a rebranding exercise and expanded to more schools. Major expansion involving more branches of art and culture, programme goals, and wider target population (primary and secondary schools) took place after a Cabinet decision in September 2006. The main objective of PBSB is to produce a society whose members can appreciate and practise cultural arts as part of their life.

Cultural Arts Guidance Program (PBSB)

The PBSB is a joint programme between JKKN and the Ministry of Education (coordinated by State Education Departments). The JKKN provides certified trainers, PBSB modules and pay trainers' salaries. Schools provide students and teacher coordinator/advisors, prepare the schedule for PBSB training (4 hours per week) as well as provide space for activities. In 2013, 733 primary and secondary schools participated in PBSB in different areas. From the list of about 400 schools given by JKKN, it was determined that the most popular branches of cultural arts were dance (227), music (94) and theatre (224). Other branches of cultural arts such as traditional games, martial arts, visual arts and language art were popular only in very few schools. However, no comprehensive study was ever conducted to evaluate the effectiveness of the implementation of PBSB until 2013. Thus, a study on the effectiveness of the programme was carried out in 2013 to help the government, particularly JKKN, to further improve programme implementation and set forth the future direction of PBSB.

PBSB Effectiveness Study

The 2013 PBSB effectiveness study was aimed at determining how effective the implementation of PBSB in schools has been as well as to examine to what extent the students involved in PBSB demonstrated achievement in aspects of a) interest in art and culture b) level of basic knowledge and skills in cultural arts c) choice of future career and d) practice of good values. The overall aim of the 2013 study was to show if there was any significant difference between primary and secondary school students in the aspects of interest in art and culture, level of basic knowledge and skills in art and culture, choice of future career and practice of good values. The study population was all primary and secondary school students who participated in PBSB. The focus of study was on the three most popular cultural arts branches – dance, music and theatre (Siti Eshah Mokshein et al., 2015).

The 2013 study used both quantitative and qualitative methods involving survey questionnaire (plus the basic knowledge test); observation of PBSB activities; and interviews with teacher advisors, PBSB trainers, parents and school administrators. For the survey, voluntary sampling technique was used. The PBSB teacher advisor or PBSB coach from each participating school was contacted by telephone and an online link of questionnaire using survey monkey was sent to them. The coach or teacher advisor then gave the link to his/her students in the PBSB group. He or she also helped the researchers to connect with three to five parents whose children participated in PBSB. Due to the low completion rate of the online survey after the first month, copies of survey questionnaires were mailed by post to the remaining schools. For the observation and interview, stratified random sampling technique was used. Thirty (30) schools were randomly chosen to represent all six zones in Malaysia – northern zone, eastern zone, central zone, southern zone, Sabah, and Sarawak.

Results of the PBSB effectiveness study showed that most schools that participated in PBSB were schools that traditionally have already been active in the specific areas of cultural arts such as music or dance. A few schools joined PBSB because of the interest of the school heads. Schools applied to JKKN through the State Education Department and JKKN then provided certified trainers to participating schools. Membership was opened to all students but limited to about twenty (20) students per group. Some schools, with high demands from their students, set academic excellence as a condition for membership. The PBSB teacher advisors were normally appointed by schools based on the former's interest or talent. Four (4) hours a week were allocated either on Wednesday or Saturday for PBSB activities. Some schools were found to conduct their PBSB activities twice a week with two hours each session. The monitoring of PBSB implementation is mostly done by the schools (principal, senior assistants, or teacher advisor) and sometimes the JKKN.

Students reported that they participated in PBSB mainly because they enjoyed PBSB activities (87.3%); some liked the interaction within the group (80.2%) while others because of their great interest in cultural arts (78.0%). It is interesting to note that similar responses came from both primary and secondary school students. Primary school students reported that the main motivators for their participation were their teachers (40%), parents (27%) and themselves (21%). The secondary school students reported the same, but the ranking and percentages differed slightly with teachers (30%) being the highest, followed by self (27%) and parents (20%). These percentages were obtained by dividing the number of respondents who checked on that particular item with the total number of respondents.

On the question about choice of future career, about two-thirds of the students chose to have a career related to cultural arts (69.3%); to pursue their education in the field of cultural arts after school (61.9%); and to become an ordinary person who can appreciate cultural arts (62.9%), which was the main aim of the PBSB. In terms of basic knowledge in cultural arts, students appeared to possess reasonable basic knowledge in the three areas of cultural arts studied. Secondary school students performed significantly better in the basic knowledge tests of music, dance and theatre. The mean differences between the two groups were about three points for dance and theatre and 16 points for music. This suggests that primary and secondary school students possess a different level of basic knowledge test in cultural arts, especially in music. Results of a one-way analysis of variance are presented in Table 1.

The effects of student participation on the development of their soft skills and personality are found to be most striking. Students reported that participation in PBSB has helped them gain basic knowledge and skills in cultural arts (89.1%); improved self-confidence (89.2%); improved self-discipline (83.5%); increased focus (80.2%); better communication skills (85.7%) and problem solving (80.4%). The PBSB also has taught them about teamwork (88.1%); sense of group belonging (84.9%) leadership (79%) and time management (77.7%).

Overall, PBSB was well implemented in the participating primary and secondary schools. The PBSB objectives have been achieved based on its benefits to students, basic knowledge and skills gained and the invitations received by the groups to perform in numerous functions. Some students who have participated in PBSB pursue their studies in cultural arts at the tertiary level in the public higher education institutions such as Universiti Teknologi MARA (UiTM), Universiti Sains Malaysia

(USM) and the National Academy of Arts, Culture and Heritage (ASWARA). School administrators, teachers, parents and students agreed that participation in PBSB brings positive effects on students, especially in the development of personality, social skills, and inculcation of basic knowledge in cultural arts. Variation between schools exists in terms of their level of commitment to implement the programme successfully, which depends a lot on the initiatives of individual schools and the additional resources that they have.

Table 1 Student Performance in Basic Knowledge Tests of Dance, Music and Theatre

Field		N	Mean Score	Std. Deviation	Mean square	F	Sig.
Theatre	Primary school	155	56.56	13.89	1142.24	4.70	.031
	Secondary school	286	59.93	16.44	243.15		
	Total	441	58.75	15.66			
Music	Primary school	192	43.23	20.64	25370.71	36.47	.000
	Secondary school	216	59.03	30.58	695.62		
	Total	408	51.59	27.50			
Dance	Primary school	549	54.46	16.19	1267.63	5.02	.025
	Secondary school	229	57.26	15.15	252.62		
	Total	778	55.29	15.93			

Research Questions

Even though the 2013 PBSB effectiveness study showed very encouraging results, the development and validation of the basic knowledge tests in dance, music, and theatre for PBSB was not given special emphasis due to the short time frame for completion of the study. Even though the test development followed the necessary procedures, data gathered was analysed in the light of classical test theory and detailed item analyses using more sophisticated tools were not performed on the data. Thus, this present study focuses on the development and revalidation of the basic knowledge test in music used in the PBSB study. This is important to ensure that the basic knowledge test administered was of high quality and psychometrically sound so that the results could be accepted with confidence.

This study mainly involved the re-analysis of data gathered to further establish validity and reliability of the instrument using Item Response Theory (IRT). Specifically, the present study attempted to address the following questions:

(i) Which IRT model fits the Basic Knowledge Test in Music for PBSB data best?

(ii) How well is the Basic Knowledge Test in Music for PBSB in terms of item fit?

(iii) How is the distribution of students' ability (θ) compared with the distribution of item difficulty (item-person map)?

(iv) How good is the basic music knowledge test for PBSB in terms of item parameter estimates?

TEST DEVELOPMENT

The Basic Knowledge Test of Music for PBSB was developed based on PBSB modules, particularly Level 1 and 2. There were four music modules especially developed by JKKN to be used by trainers in the implementation of PBSB. Each module contains different materials, with Level 1 being the introductory level and Level 4 the more advanced level. Two major components of the modules are theory (which includes history and appreciation) and technical skills. The Level 4 Module, however, focuses only on the development of technical skills among the participants.

Music Modules

The theory component of Level 1 Music Module consists of two parts, namely the basic music theory and appreciation of music. The main objective of this component is to enable students to write musical notes clearly and differentiate note values accurately. It will also help students to understand the use of musical notes and musical terms and read simple rhythms. Course content covers writing staff, key signatures, rest signature, treble and bass clef, shape and note values, writing scales, formation of triad and simple musical terms.

The musical appreciation component exposes students to classification of musical instruments, shapes, structures and concepts and ideas in songs. The history of the music of Malaysia is also being introduced. The main purpose is for students to understand and appreciate the aesthetic values of the traditional music of Malaysia and be able to differentiate traditional and modern music. Students will also be introduced to the function and roles of music in the context of dance, play, ensemble, vocal and instrumental music. Several types of songs introduced are traditional, ethnic, folk, and contemporary songs. The technical skills introduced in the Level 1 Music Module include playing '*paluan kompang*', basic '*paluan gendang muzik silat Kedah*', '*paluan marwas*, *paluan rebana Melayu*', '*cak lempong*' and basic '*gamelan'*. In the Level 2 Music Module, the basic technical skills learned in Level 1 Module are further developed.

The Level 2 Music Module emphasises western music techniques as well as traditional musical rudiments. The objective is to enable students to write and differentiate note values accurately and understand the concept of intervals. Students will also be able to read simple rhythms fluently, write scales using correct musical notations and understand the use of dynamics, key signatures and musical terms. Course content covers bass clef, name and keys, form and compound notes, time signature 6/8, key signature of 1 to 3 flats and 1 to 3 sharps, major and minor scales, number notation, history and mnemonics of Malay drumming, timbre, musical intervals and triplet.

The Basic Knowledge Test of Music for PBSB is however, confined to only the basic music theory and not other aspects of the module due to time constraints of the study and differences in emphasis of implementation by different trainers. A two-day workshop on test development was conducted in the first week of April 2013 involving four groups of test developers focusing on different aspect of PBSB effectiveness study – music, dance, theatre, and survey questionnaire to gauge participants' interest and choice of future career among others. Workshop participants were exposed to the study objectives and the PBSB modules. Three experts from the Music and Performing Arts Faculty, UPSI were invited to develop the music test. A table of content was prepared based on the modules to establish content validity and draft of items were then written to make the test. The draft of the test was brought back to the faculty to be further vetted and reviewed. Two weeks later, the draft of Basic Knowledge Test in Music for PBSB was presented to the research team and it was accepted with some recommendation.

In May 2013, the research team then tried out the music test with 20 PBSB participants from a primary school in Batang Padang District, Perak to find out the suitability of language and terms used. The test was further refined based on the issues encountered by participants while attempting the questions and also the views of the trainer. The test was then finalised and used in the effectiveness study in July 2013. The internal consistency of the test as measured by Cronbach alpha was 0.884. However, further item analysis was not performed on data as the major focus of the PBSB effectiveness study was to determine the overall effects from various perspectives and the results were needed urgently. Therefore, this present study focuses on the item analysis and revalidation of the test using dichotomous Item Response Theory (IRT). Two files, namely data matrix file and data control file, were created from the dataset in SPSS format before 1PL, 2 PL and 3PL analyses were performed on the data using Xcalibre 4.2.

Why Item Response Theory (IRT)?

Even though the Classical Test Theory (CTT) is widely used by most educators at all levels, several issues regarding CTT may affect the precision of the measurement and subsequent analysis. Firstly, the number of correct responses or raw scores determines ability. In reality, a test represents only a sample of items measuring specific objectives. If different samples of items are administered, students will likely to obtain different scores each time. Secondly, test difficulty is dependent on the test takers. If a test is administered to different groups of people, different difficulty values will be obtained. The same test administered to high ability students for instance, will yield lower difficulty level compared with the test administered to low ability students. Thirdly, students obtaining similar scores are assumed to possess similar ability level, regardless of the difficulty of the items that they had answered correctly. Item difficulty is not taken into account in determining the ability of the person. Thus, in CCT, the difficulty of the test is dependent on the person's ability and the ability of the person is dependent on the test difficulty.

The Item Response Theory (IRT) addresses the issues highlighted above successfully. Unlike the Classical Test Theory, in which the test scores of the same examinees may vary from test to test depending on the test difficulty, item parameter calibration is sample-free while examinee proficiency estimation is item-independent in

IRT (Chong, 2013). The IRT expresses the relationship between an individual's response to an item and the underlying latent trait, also called construct or ability or proficiency. It is a probabilistic model whereby the probability of a person getting a correct answer for a particular item is a function of his or her ability and item parameters (difficulty, discrimination, guessing etc.).

The IRT is widely used in scoring tests and surveys and also in computer adaptive testing (CAT). The IRT scoring takes into account the item difficulty and discrimination. Items that are more discriminating, or more reliable, are weighted more heavily, making IRT scores more reliable than number-correct scores. If different examinees take different tests, the IRT scores adjust for the differences in difficulty (DeMars, 2010). Additionally, IRT can be used in test or scale development. The IRT analysis supplies indices of item difficulty and discrimination. Knowing the item difficulty is useful when building tests to match the trait levels of a target population. For example, the items on a fourth grade science test should not be so easy that the average fourth-grader answers nearly all the items correctly, nor should they be so difficult that the average student answers nearly all of them incorrectly. Similarly, an instrument intended to measure the wellbeing of a college population should not consist of items endorsed only by those with clinical depression. Another item index, discrimination, is useful for selecting items that differentiate well between examinees with low and high levels of the proficiency or attitude measured by the test items. Together, difficulty and discrimination can be used to calculate the standard error of measurement or reliability of the scores (ibid.). The units of the ability scale, called logits, typically range from -4 to 4. They represent the natural logarithm of the odds for success on the test items. For example, if a person succeeds on 80 per cent and fails on 20 per cent of the test items, the odds ratio for the success on the test is 4/1 = 4. Thus, the ability score of this person is the natural logarithm of 4 (or ln 4), which is 1.39 (Dimitrov & Shelestak, 2003).

Hambleton, Swaminathan and Rogers (1991) stated that there are three IRT models commonly used for dichotomous items, namely the one-parameter logistic model (1PL model), the two-parameter logistic model (2 PL model) and the three parameter logistic model (3PL model), so named because of the number of item parameter each model incorporates. As the number of parameters in the model increases (for example, from 1 to 2 to 3), the model becomes more flexible and thus, can provide a more realistic reflection of how the expected response to each item is related to the underlying ability. The c parameter (or guessing parameter) is the probability of a candidate with very low ability to get a correct response on the item. DeMars (2010) argued that even someone who does not have knowledge about the subject has a chance to get a correct response in multiple choice items. Meyer & Shin-Zu (2013) stated that 3PL is the most common model for dichotomous items. The mathematical models for 1PL, 2PL and 3PL are shown in equations (1) – (3).

$$\text{1PL:} \quad P_i(\theta) = c_i + (1 - c_i)\,[\,1 + e^{-Da(\theta - bi)}]^{-1} \quad \text{.......} \quad (1)$$

$$\text{2PL:} \quad P_i(\theta) = c_i + (1 - c_i)\,[\,1 + e^{-Da(\theta - bi)}]^{-1} \quad \text{.......} \quad (2)$$

$$\text{3PL:} \quad P_i(\theta) = c_i + (1 - c_i)\,[\,1 + e^{-Da(\theta - bi)}]^{-1} \quad \text{-------} \quad (3)$$

where;

P_i (θ) = the probability that a candidate with ability theta (θ) will answer item i correctly,
b_i = difficulty parameter for item i;
a_i = discrimination parameter for item i (in 1PL a = constant),
c_i = guessing parameter for item i (in 1PL and 2PL models = constant)
n = number of items in the test
D = scale factor (D=1.72)

Two important assumptions in IRT are unidimensionality and local independence. Unidimensionality means that only one single latent factor is measured (ability/ proficiency), whereas local independence means that individual response on an item does not depend on his/ her response on other items. Local independence will be obtained if unidimensionality is met (Lord, 1980; Lord & Novick, 1968 in Hambleton et al., 1991). Thus, exploratory factor analysis using SPSS was performed on the music data in this present study to test whether the use of IRT analysis was appropriate for the data.

Unidimensionality assumption also means that for a set of items in a test, each person has only one theta value. Three factors were extracted in the exploratory factor analysis on the data with factor 1 contributing to 32.9% to the total variance explained. This is sufficient as according to Reckase (1979), more than 20% variance explained is needed for accurate estimation (Reckase, 1979). The number of component extracted and variance explained are shown in Table 2.

Table 2 Components extracted and total variance explained for the Basic Knowledge Test of Music for PBSB (BKToM-PBSB)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.606	32.901	32.901	4.606	32.901	32.901
2	1.609	11.492	44.393	1.609	11.492	44.393
3	1.150	8.216	52.609	1.150	8.216	52.609
4	.947	6.764	59.373			
5	.922	6.586	65.959			
6	.780	5.571	71.530			
7	.603	4.306	80.706			
8	.571	4.077	84.783			
9	.509	3.635	88.418			
10	.476	3.402	91.819			
11	.458	3.274	95.094			
12	.379	2.710	97.804			
13	.307	2.196	100.000			

Analysis of scree plot shows a big jump, indicating that possibly there is only one dominant factor present in the test (de Ayala & Hertzog, 1991). Thus, the unidimensionality and local independence are assumed and IRT analyses can be performed on the music data. The scree plot obtained for the music data is shown in Figure 1.

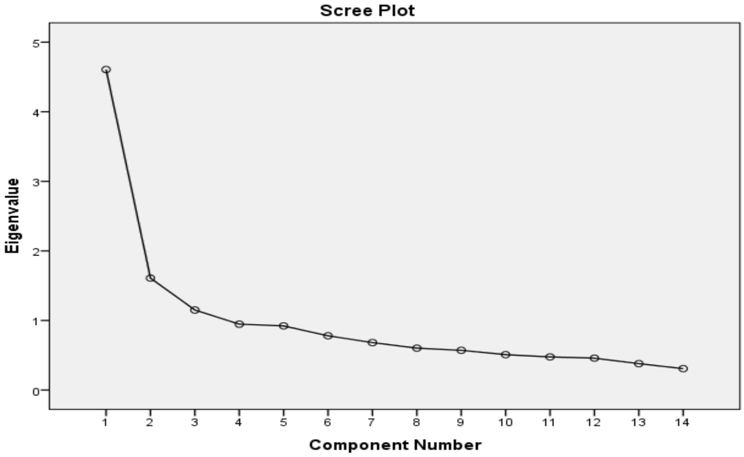


Figure 1 Scree plot of EFA for the Basic Knowledge Test of Music for PBSB

RESULTS AND DISCUSSIONS

Which Item Response Theory Model?

All the three parameter logistic models (1PL, 2PL and 3 PL) have the potential to be used for the multiple choice test of Basic Knowledge in Music for PBSB. Which model is most appropriate for the data, however, depends on several considerations such as overall fit, comparison of -2*log likelihood (Thissen, 1991), graphical representation of fit (Hambleton & Swaminathan, 1985) and distribution of chi-square statistics, z-residual and other parameters (such as difficulty parameter, b and discrimination parameter, a). The Xcalibre outputs from the three models as shown by the overall fit, parameters and theta estimates, -2*log likelihood (-2LL), and graphical representation of fit suggest that the two-parameter model (2PL) is the best model for the music data. The 2PL model yielded the lowest -2LL and chi-square values. It also produced the most stable distribution of theta estimates. In terms of item misfit, both 2PL and 3PL models produced one item misfit (Item 3), whereas the 1PL model yielded six item misfit (Items 1-3; and 9, 10, 13). Further analysis showed that these are easiest (Items 1-3) and hardest items (9, 10, 13). Details of the output are presented in Table 3.

Table 3 Comparison of output for 1PL, 2PL and 3PL models

a) Overall Model Fit

Model	Item	*Chi-square*	*df*	*p*	*-2LL*
1PL	14	1033.208	196	0.00	6165
2PL	**14**	**567.386**	**182**	**0.00**	**5557**
3PL	14	581.209	168	0.00	5994

b) Mean and SE for Theta and Item parameters

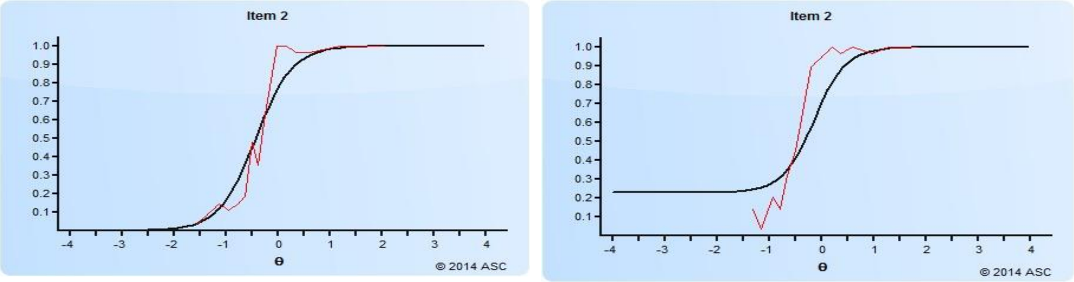
Model	Parameter	Mean	SD	Min	Max
1PL	Theta	0.072	1.094	-2.120	2.126
	b	0.00	1.00	-1.675	1.849
2PL	**Theta**	**0.00**	**1.00**	**-7.000**	**7.000**
	b	-0.021	0.455	-0.793	0.844
	a	1.016	0.417	0.512	1.846
3PL	Theta	0.024	1.026	-1.400	1.830
	b	0.552	0.635	-0.275	1.577
	a	2.068	0.691	1.000	3.238
	c	0.240	0.024	0.208	0.292

c) Item Misfit

Model	No of Items	Item	Flag	
1PL	6	1, 2, 3, 9,10,13	F	(Easiest and hardest items)
2PL	1	3	F	
3PL	1	3	F	

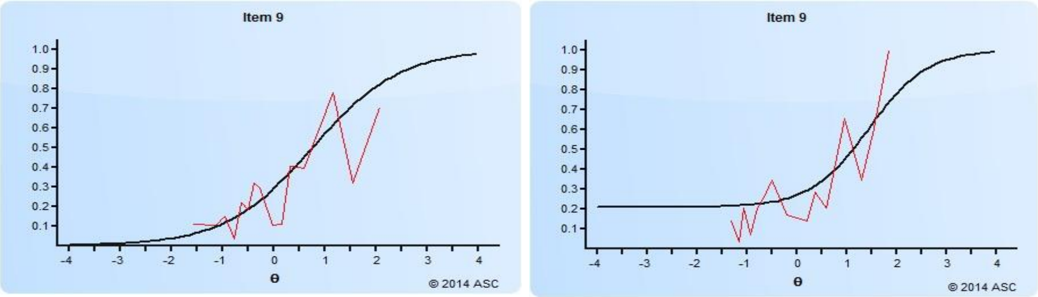
The 2PL and 3PL model output also showed similarities in the pattern of item parameters. Even though the values differ, the order of item difficulty and item discrimination holds the same for both models. Thus, the Item Characteristic Curves (ICC's) of the most discriminating item and the hardest item of 2PL and 3PL models were explored to demonstrate the graphical representation of goodness-of-fit between the two models. The ICCs for the items are shown in Figures 2 and 3.

The ICC's of the two items (Items 2 and 9) showed that the 2PL model fit the music data better compared with the 3PL model. Similarly, examination of test information function (TIF) of both models also showed that more information is yielded from the 2PL model (Figure 4). Accordingly, the cumulative standard error of measurement (CSEM), an inverted function of the TIF, which estimates the amount of error in theta estimation for each level of theta was smaller for 2PL model. Thus, 2PL model fits the Basic Knowledge Test of Music for PBSB (BKToM-PBSB) the best.



2PL Model 3PL Model

Figure 2 ICCs of the most discriminating item



2PL Model 3PL Model

Figure 3 ICCs of the hardest item

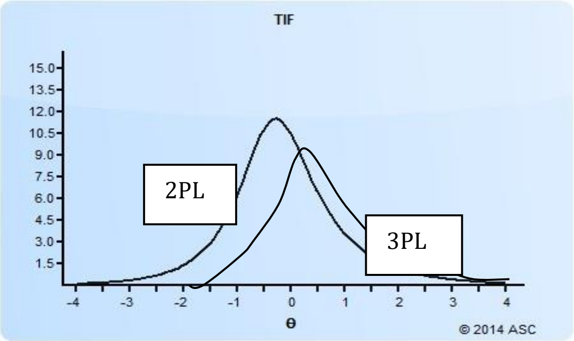


Figure 4 Test information function (TIF) of the BKToM-PBSB

Theta and Item Parameter Estimates

Analysis of 2PL model showed that the theta estimates of the 437 PBSB students range from -7 to +7.0 with the mean 0.00 and standard deviation equals 1.0 (Table 4). Theta estimates for all calibrated items are represented in Figure 5.

Table 4 Summary statistics for the theta estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max
Full Test	437	0.000	1.000	0.587	-7.000	-0.721	-0.196	0.750	7.000

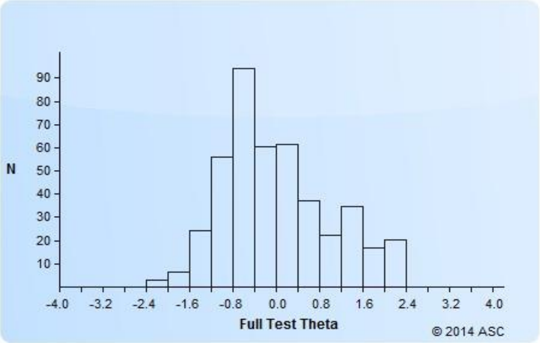


Figure 5 Theta estimates for all calibrated items

The mean for the difficulty parameter of the items, b, was -0.021, slightly lower than the mean ability. The mean discrimination parameter for the items, a, was 1.016 (Table 5).

Table 5 Summary statistics for all calibrated items

Parameter	Items	Mean	SD	Min	Max
a	14	1.016	0.417	0.512	1.846
b	14	-0.021	0.465	-0.793	0.844

Item ID	P	R	a	b	Flag(s)
1	0.691	0.413	0.801	-0.793	
2	0.606	0.613	**1.635**	-0.365	
3	0.595	0.632	**1.846**	-0.330	F
4	0.506	0.432	0.840	-0.056	
5	0.556	0.637	**1.582**	-0.224	
6	0.554	0.511	1.055	-0.220	
7	0.616	0.517	1.142	-0.419	
8	0.446	0.587	**1.194**	0.121	
9	0.307	0.366	0.678	**0.844**	
10	0.373	0.255	0.512	**0.654**	
11	0.414	0.493	0.859	0.287	
12	0.458	0.464	0.789	0.133	
13	0.398	0.280	0.541	**0.494**	
14	0.596	0.407	0.758	-0.414	

Item-person map

The item-person map shows the distribution of item difficulty and persons' ability on the same scale. The map shows that the test difficulty matches the ability of most of the students. However, the test cannot provide any information about the very low ability students (theta below -0.8) and the very high ability students (theta above 1.6) in music (Figure 6).

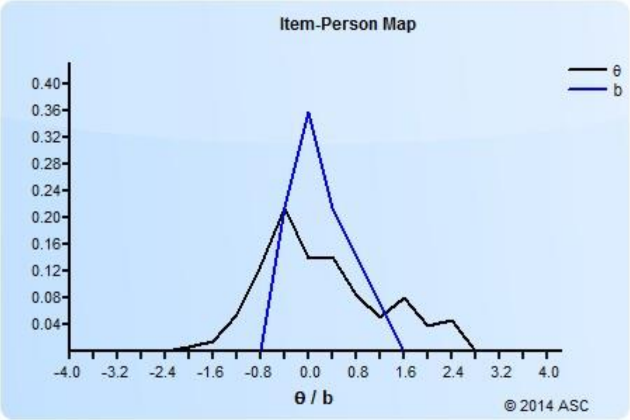


Figure 6 Item-person map

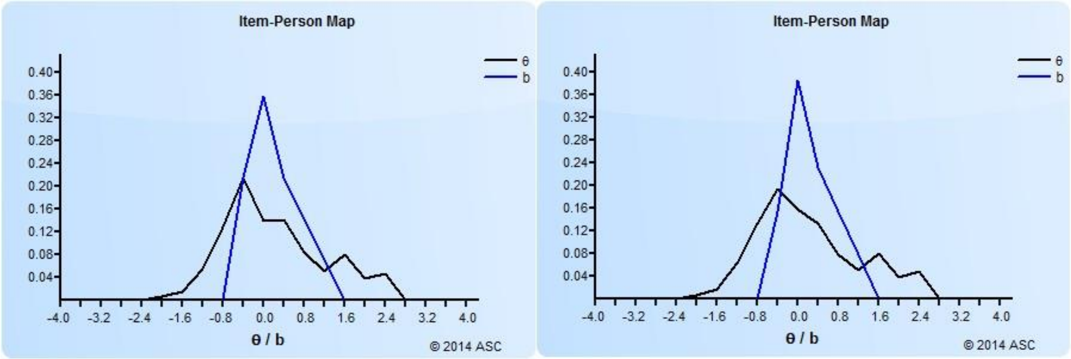
The test also shows differential item functioning (DIF) as demonstrated in the subgroup statistics (Table 6). The mean theta for secondary school students was much higher (0.261) compared with the mean theta of primary school students (-0.275). Detailed examination of the items did not suggest any element of bias in the questions. Thus, the different group means could be due to the differences in the level of basic music knowledge possessed by the two groups of students.

Table 6 Subgroup statistics for the full test

Subgroup	Examinees	Mean Theta	SD Theta
primary	201	-0.275	0.686
secondary	212	0.261	1.169

DISCUSSION

Results of the analyses showed that all the items in the Basic Knowledge Test of Music for PBSB possess good psychometric characteristics except for Item 3 which has a 'F' flag, indicating that this item did not fit the model. Further examination of the item-person maps, however, did not show much difference whether or not Item 3 is included (Figure 7).



Full test of 14 Items

Test with Item 3 deleted

Figure 7 Comparison of item-person maps

Comparison of item parameters with and without Item 3 included also did not show much difference (Table 7). In addition, Item 3 showed to have acceptable values of difficulty (b) and discrimination (a) parameters. Furthermore, the difficulty and discrimination parameters of the items also did not differ much whether or not Item 3 was included. This indicates that the exclusion of Item 3 did not improve the precision of the parameter estimates. In addition to the above, the Item Characteristic Curves (ICC) of Item 3 was compared with the ICCs of several other items such as Item 2, Item 6 and Item 7. The ICC of Item 3 is quite similar to that of Item 3 (Figure 8).

Table 7 Comparison of item parameters with and without Item 3 included

	2PL 14 Items				2PL 13 Items			
Item ID	P	R	a	b	P	R	A	b
1	0.691	0.413	0.801	-0.793	0.691	0.413	0.837	-0.765
2	0.606	0.613	**1.635**	-0.365	0.606	0.576	**1.414**	-0.362
3	0.595	0.632	**1.846**	-0.330				
4	0.506	0.432	0.840	-0.056	0.506	0.412	0.778	-0.045
5	0.556	0.637	**1.582**	-0.224	0.556	0.623	**1.617**	-0.214
6	0.554	0.511	1.055	-0.220	0.554	0.492	1.025	-0.212
7	0.616	0.517	1.142	-0.419	0.616	0.499	1.127	-0.410
8	0.446	0.587	**1.194**	0.121	0.446	0.578	**1.196**	0.130
9	0.307	0.366	0.678	**0.844**	0.307	0.372	0.699	**0.837**
10	0.373	0.255	0.512	**0.654**	0.373	0.259	0.517	**0.659**
11	0.414	0.493	0.859	0.287	0.414	0.503	0.895	0.287
12	0.458	0.464	0.789	0.133	0.458	0.472	0.823	0.137
13	0.398	0.280	0.541	**0.494**	0.398	0.277	0.546	**0.500**
14	0.596	0.407	0.758	-0.414	0.596	0.408	0.792	-0.396

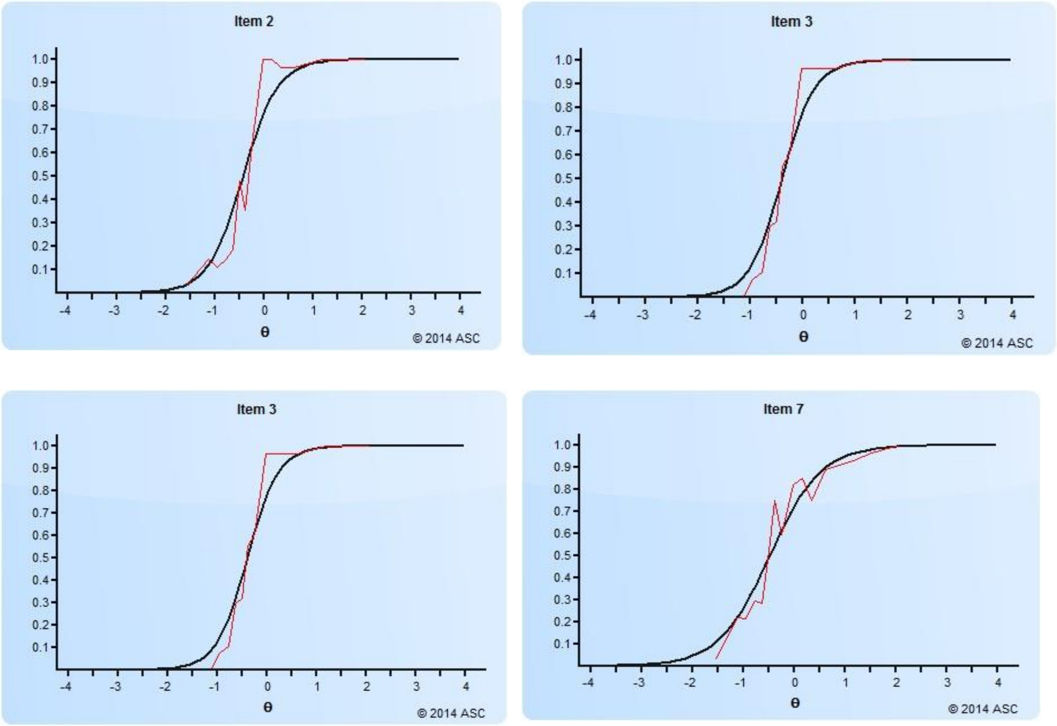


Figure 8 Comparison of Item 3’s ICC with ICCs of other items

Since the parameter estimates of Item 3 were found to be sound, the ICC of the item looked similar to ICC of other items that fit the model and the exclusion of the item did not improve the precision of TIF and parameter estimates of other items. As the test contained limited number of items, the researchers decided to include Item 3 in the test.

CONCLUSION AND RECOMMENDATION

The Basic Knowledge Test of Music for PBSB (BKToM-PBSB) was found to possess good psychometric characteristics as reflected by the model fit, the item-person map, reliability and validity of ability estimates, and the difficulty and discrimination parameters. The item-person map showed major overlap between the item difficulty and the students’ ability, indicating that for most students, the test difficulty matched their ability. However, the test was too difficult for very low ability students and too easy for the advanced level group in terms of basic music knowledge. This was anticipated as students who participated in the study were expected to possess different levels of music knowledge depending on the level of modules that they have completed. The means ability for the two sub-groups (primary and secondary school students) were noticeably different, with that of primary school students being lower than the secondary school students. Since no elements of bias could be detected in the items, this suggests that most students in primary school were at lower-level module while their counterparts in secondary schools were mostly at more advanced level during the time of the test. It is recommended that different tests be developed for different levels of modules that the students undertake if BKToM-PBSB is to be used to complement the existing PBSB assessment systems. Currently, some forms of assessments are conducted by the individual trainers at the end of each level but the uniformity of the assessment is not known and the results are not readily accessible, causing nationwide evaluation very difficult.

REFERENCES

Chong, H. Y. (2013). *A simple guide to the Item Response Theory (IRT) and Rasch modeling.* Retrieved from http://www.creative-wisdom.com/computer/sas/IRT.pdf, downloaded on 2 March 2015.

de Ayala, R. J., & Hertzog, M. A. (1991). The assessment of dimensionality for use in item response theory. *Multivariate Behavioral Research, 26,* 765-792.

DeMars, C. (2010). *Item Response Theory: Understanding statistic measurement.* New York, NY: Oxford University Press, Inc.

Dimitrov, D. M., & Shelestak, D. (2003). Psychometric analysis of performance on categories of client needs and nursing process With the NLN Diagnostic. *Journal of Nursing Measurement, 11* (3), 207-223.

*Hambleton, R. K., Swaminathan, H., & R*ogers, H. J. (1991). *Fundamentals of Item Response Theory*. USA: Sage Publications, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Lord, F. M. (1980). *Application of Item Response Theory to practical testing problem*. Hillsdale, New Jersey: L Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley.

Meyer, P. J., & Shi-Zhu. (2013). Fair and equitable measurement of student Learning in MOOCs: An introduction to Item Response Theory, scale linking, and score equating. *Research & Practice in Assessment, 8,* 26-39.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Siti Eshah Mokshein et. al. (2015). *Penilaian keberkesanan program bimbingan seni budaya (PBSB) di sekolah-sekolah Malaysia [Evaluation of the effectiveness of the cultural arts program (PBSB) in Malaysian schools]*. Malaysia: UPSI Publisher.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, Categorical item analysis and test scoring using item response theory.* Chicago: Scientific Software.

BIOGRAPHY

Siti Eshah Mokshein is an associate professor (Educational Measurement & Evaluation) in the Faculty of Education, Sultan Idris Education University since 2008. Prior to joining the university, she spent many years of her career in the Ministry of Education Malaysia serving the Federal School Inspectorate and the Educational Policy Planning and Research Division. She obtained her PhD in Education from the University of Iowa, USA.
Email: eshah@fppm.upsi.edu.my

Zaharul Lailiddin Saidon is an associate professor of Music Education at the Sultan Idris Education University (UPSI), Tanjong Malim, Malaysia, where he teaches in music education and marching band techniques. Zaharul currently serves as the Dean of the Faculty of Music and Performing Arts at the university. He received his undergraduate degree in music education at Southern Illinois University, and his master's degree from University of Houston, Texas. Zaharul is the founding member of the Malaysia Band Association and the Malaysian Association for Music Education.
Email: zaharul@fmsp.upsi.edu.my

Brian Doig is a senior lecturer in the Department of Mathematics, Faculty of Arts and Education, Deakin University, Australia. He is an experienced survey developer and analyst. He has been involved in international assessment studies, such as TIMSS and PISA. His research interest include the use of the Rausch model for the analysis of ordinal data form surveys and interviews.
Email: b.doig@deakin.edu.au