ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

Assessing Psychometric Properties of Malaysian Secondary School Students' Arabic Vocabulary Knowledge Inventory

Zunita Mohamad Maskor
Faculty of Education, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia
Tel: +6012-2511944 E-mail: kobiscomel73@gmail.com

Harun Baharudin
Faculty of Education, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia
Tel: +6012-3782207 E-mail: harunbaharudin5291@gmail.com

ABSTRACT

This study examines the use of psychometric properties to measure the Arabic vocabulary knowledge of Islamic secondary school students in Malaysia. An instrument was developed as a diagnostic test to measure various dimensions of Arabic vocabulary knowledge based on the Rasch Model. The study may give insight into the depth of Malaysian secondary school students' understanding of Arabic language in word meaning. This paper also discusses the development of the Arabic Vocabulary Knowledge Inventory (PekkA). The instrument consisted of 57 items of vocabulary knowledge that focused on the aspects of meaning namely denotative meaning, semantic relationship, word formation, and word combination. 483 respondents were randomly selected from 39 Islamic secondary schools in Peninsular Malaysia. The study employed a cross-sectional quantitative research design. The data was analyzed using the WINSTEP, a Rasch Measurement Model software to determine the validity and reliability of the instrument. The result of the study confirmed the psychometric requirement by Rasch model with good reliability and validity. In general, the results suggest that the instrument could be used as a reliable research instrument for evaluating the vocabulary knowledge among Arabic language learners in secondary school settings.

Keywords: Vocabulary knowledge, Validity, Reliability, Arabic Vocabulary Knowledge Inventory (PekkA), Rasch Measurement Model.

INTRODUCTION

Vocabulary knowledge is defined as the ability to learn words that have a profound meaning, including pronunciation, meaning, spelling, frequency, sound structure, syntax and collocation by context (Qian, 2002). Meanwhile, Haastrup and Henriksen (2000) defined vocabulary knowledge as word knowledge from the point of view of meaning, cognition, and collocation. From a broader perspective, Nation (2001) saw the development of vocabulary knowledge by combining form (pronunciation, spelling, and words), meaning (structure or meaning of words, ideas, and preferences or a combination of words) and use (syntax, collocation, constraints).

Arabic language, knowledge of vocabulary emphasizes knowledge of the meaning of the word in depth. The Arabic language learners should know the translation and synonymous word with target words in terms of meaning (Al-Shuwairekh, 2001). In terms of use, learners need to know how the word is used as a phrase (Harun, 2014). Thus, knowledge of the meaning may increase the students' ability to connect one word with another word using accurate syntax. This is no different to the definitions of vocabulary knowledge in other foreign languages. However, knowing a word is useless unless the meaning of the word is clearly understood. Most language researchers perceived vocabulary knowledge as a multidimensional construct in word knowledge. They recommended that knowing

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

a word ought to incorporate different kinds of linguistic knowledge from pronunciation, spelling, morphology, syntax, and semantic relationship including synonym, antonym, hyponym, polysemy and collocation (Nation, 2001; Qian, 2002; Milton, 2009).

Words in Arabic are mostly based on sound, spelling and their specific meaning. An Arabic word is flexible to any addition, elimination and replacement of letters that indicate a particular function and meaning (Che Radiah & Norhayuza, 2013). As debated by Henriksen (1999), Nation (2001), Qian (2002) and Milton (2009) regarding vocabulary knowledge, Al-Shuwairekh (2001) categorized Arabic vocabulary knowledge into four aspects namely 1) knowing the words and it patterns, 2) applying morphology in words to form different words, 3) distinguishing oral word with written word, and 4) concluding a short pronunciation within context. In accordance to this, Al-Naqah (1985) explained that knowing a word is an ability to combine the Arabic letters (known as *harf hijaiyyah*) until they form an understandable word orally or in written text.

Arabic Vocabulary Learning in Malaysian Secondary Schools

Unlike their counterparts, the native speakers from Arabic countries, the approach to language learning practiced in Malaysia might differ significantly. As non-native speakers of the Arabic language, Malaysian secondary school students emphasized the acquisition of vocabulary as a vital indicator of competency in the language. Conventionally, most students acquired new Arabic words through the recitation method. The words acquired are useful both in oral or written sentences. However, they will soon experience difficulties as they realized they lack the knowledge of words meaning due to the Arabic language learning in the secondary school level has more emphasization in the translation of the text (Maskor, Baharudin, Lubis & Yusuf, 2016).

In order to improve their language skills, students will have to learn vocabulary themselves. It is important for the educators to guide their students in language activities to ensure the students are on the right track as vocabulary enhancement is achievable through a variety of learning approaches. The educator's role should not only be limited to translating texts but also to involve students in finding new words in the text. Students will be able to acquire vocabulary effectively through this interaction.

Mastering vocabulary means a student is able to pronounce words with the right *makhraj*, accurately grasp the meaning of words, find out the origin of their fractions and derivatives, and to understand the proper way to arrange them in full sentences. The main outcome expected from the activities is that the students will be able to choose the right vocabulary for its meaning. Students may find it challenging as they are unable to discern the meaning of a word by its grammatical function, the combination of words, word formation and use (Tha`imah, 1989).

Chapelle (1994) also stated the non-native speakers can acquire a number of words but will fail to use them because of their weakness in understanding word meaning. This is in line with Schmitt's (2000), Nation (2001) and Che Radiah & Norhayuza (2013) where they pointed out that the vocabulary knowledge is a complex aspect of language that covers the aspects of form, meaning, purpose and function of the word.

The above said situation clearly demonstrates that acquisition of vocabulary is geared towards its use in language skills as demanded in the curriculum. Therefore, it is necessary to develop a comprehensive instrument to assess the level of the vocabulary acquisition among secondary school students in Malaysia. Teachers can assess the vocabulary skills of their students in detail with the appropriate instrument. With this in mind, we believe teachers can identify teaching activities that correspond to the level of vocabulary.

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

The Arabic Vocabulary Knowledge Inventory (PekkA)

In the foreign language study, vocabulary knowledge tests were developed to assess the depth of learners' knowledge of words without regard to their learning background. There are many vocabulary knowledge tests used to assess foreign language learners' vocabulary knowledge such as the Vocabulary Knowledge Scale or known as VKS (Wesche & Paribakht, 1996), the In-depth Vocabulary Knowledge or known as DVK (Qian & Schedl, 2004), the Word Association Format (Schmitt, 2010) and many more.

However, most of these tests were developed in the western countries for the English language, and there is a limited number of tests for vocabulary knowledge in the Arabic language for Malaysian learners. Examples of the Arabic vocabulary test that had been developed were TAV or Test of Arabic Vocabulary (Harun, Zawawi, Adelina & Normala, 2014) and Arabic Vocabulary Level Test or aVLT (Abdul Razif & Mohd Zaki, 2015). However, the tests have more focus on vocabulary size instead of vocabulary depth. Hence, there is a need for a valid and reliable test to measure vocabulary knowledge in the Arabic language apart from standard school examination papers.

The instrument was developed based on the theoretical framework of Bachman (1990) and Chapelle (1994) who defined vocabulary ability as the combination of language knowledge and the ability to use language in context. In this case, Chapelle (1994) gave a more comprehensive definition of vocabulary ability, known as the context of vocabulary use, vocabulary knowledge and fundamental processes, and metacognitive strategies for vocabulary use. In accordance with the theoretical framework, the Arabic Vocabulary Knowledge Inventory (PekkA) aims to assess students' knowledge of different aspects of words with focus on meaning aspect (Alderson, 2005; Nation, 2001). The meaning aspects then were divided into denotative meaning, semantic relations, words formation and word combination (Alderson, 2005). This is based on the vocabulary test developed by Alderson (2005), which applied the Common European Framework of Reference standards (CEFR). These word meaning aspects were defined as the ability to recognize or produce word meanings, including denotative meaning, semantics relationship, synonym, antonym, hyponym, hypernym, polysemy, collocation, idiomaticity, word compounding and affixation (Alderson, 2005).

In order to utilize a sampling approach, the words were chosen based on the frequency list from Buckwalter and Parkinson (2011), which used 1500 words as listed in the syllabus. Then, the selected words were refined using the words listed in the Arabic Language Syllabus (2006). If the words selected from Buckwalter and Parkinson (2011) were not included in the Arabic Language Syllabus from Ministry of Education Malaysia (2006), the word would be omitted and replaced with other words found in both sources. The items were developed by adopting words from Form One to Form Four Arabic language textbooks. As such, most of the words were considered familiar to the respondents because the textbooks were widely used as it is compulsory learning aids for Arabic language lesson in all Islamic secondary schools under the supervision of the Ministry of Education. The items difficulties classifications were estimated based on nine experts' reviews and authors' own experience in teaching the Arabic language.

Psychometric Properties in the Rasch Measurement Model

In Malaysia, most instruments were developed using Classical Test Theory (CTT) as the underlying theoretical perspective that focused on the item difficulties and item discrimination (Siti Rahayah, 2008). Item Response Theory (IRT), also known as latent trait theory is another contemporary alternative to the CTT

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

in social science measurement. IRT approach emphasizes on the probability of the responses given by the respondents, which are influenced by the level of individual abilities (latent trait) and test items difficulty (Embretson & Reise, 2000; Siti Rahayah, 2008).

IRT perspective helps to recognize the components that affect the likelihood of the way a respondent responds to an item. A measurement model communicates the mathematical associations between an outcome and the components that influence the outcome. The outcome comprises of a respondent's score on an item. Meanwhile, the components are the qualities that are portrayed by the respondent or the item. The significant difference between the measurement models is in the terms of item characteristics or parameters, and the response option format. IRT is the only test theory that has the characteristics of the scaling properties of linear interval measurement (Gorin & Embretson, 2008; Embretson & Reise, 2000).

Consequently, the Rasch model was chosen because of the postulations met the IRT model assumptions. According to Gorin and Embretson (2008) and Siti Rahayah (2008), the first assumption is dimensionality that infers the latent traits to be modeled entirely portray the manner underlying the item response. The second assumption is local independence between item responses that require the responses for one item are extraneous to responses for another item.

Finally, the IRT models assume a particular relationship between changes in trait level and changes in the probability of a given response which portray in graphic, known as the item characteristic curve (ICC) (McCreary, Conrad, Conrad, Scott, Funk & Dennis, 2013). However, there is no uniformity in the Rasch model evaluation sequence. Several researchers started with unidimensionality (McCreary et al., 2012), meanwhile few researchers ignore local independence and DIF. This has led us to conclude that the Rasch model evaluation varied due to the objectives of the study conducted. Therefore, this study takes a new look in depth of unidimensionality, fit statistic, local independence, and DIF as psychometric assumptions that need to comply with the Rasch Measurement Model.

METHODOLOGY

Participants

A cross-sectional quantitative survey method was employed in the study. The instrument was administered to 483 students from eight Islamic secondary schools under the supervision of the Ministry of Education. Stratified sampling technique was used to distribute samples according to states and gender. The schools were located in four main regions of Peninsular Malaysia, which are the north, east, central and south regions. The respondents were selected among Form Four students, which consisted of 143 male students (33.3%), and 340 female students (77.6%). The proportion of the respondents is equal to 483 students out of 7353 Form Four students of religious secondary schools under the Ministry of Education. The study was conducted in September 2017.

Instrumentation

The Arabic Vocabulary Knowledge Inventory (PekkA) was developed to measure secondary school students' vocabulary acquisition in the Arabic language. The test contains 20 items of dichotomous responses (True and False), 17 items of multiple-choice responses and 20 items of sentences completion response. The PekkA contains five demographic profiles of the respondents (i.e. school location, region, gender, Arabic achievement in Form Three Assessment (PT3) and Primary School Evaluation Test (UPSR).

There are three parts of the test in one booklet namely Part 1: Vocabulary Size (True-False response), Part 2: Vocabulary Depth (Multiple-choice response) and Part 3: (Sentences-completion response). The true-false response

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

item consisted of 20 items that measure the vocabulary size. While the 20 items of multiple-choice response and the 17 items of sentences completion are intended to measure the depth of the vocabulary.

These 37 items of vocabulary depth addressed four dimensions of word meanings such as denotative meanings (12 items), semantic relationships (11 items), word formation (10 items) and a combination of words (4 items) (Alderson, 2005). Before the actual data collection, a small pilot test was carried out in a religious secondary school in Perak with 102 students.

During the pilot test, the study involved a pre-testing instrument containing 30 items of vocabulary size and 50 items of vocabulary depth. The aim of the pilot study was to confirm that wording, formatting, and layout were appropriate. In the main study, those 150 students were excluded. The results of the pilot study using the Rasch model indicated that the instrument is in good reliability and validity (Maskor, Baharudin & Lubis, 2018).

Data Analysis

The data was collected and analysed using Microsoft Excel and Rasch software on the basis of partial credit model data. After the data screening, WINSTEP version 3.73 was used to determine the instrument's validity and reliability. In order to evaluate the vocabulary knowledge of students in Arabic, descriptive statistics, including mean and standard deviation were used. The mean score was in the scale of the logit which was transformed from the raw data. The bigger the logit score in terms of items meant that the student had a better ability to recognize and use vocabulary in the Arabic language.

RESULTS AND DISCUSSION

In this study, the Rasch measurement model outlined the foundation for determining the measurement quality of PekkA. WINSTEP software mathematically calibrating the item difficulties and person abilities into logit measures while concurrently evaluating the instrument fitness. The iterative process produces imperative output tables that present as a diagnostic evaluation of PekkA and showing statistical report of psychometric properties of the instrument.

Unidimensionality

Unidimensionality is a primary requirement in construct validity to verify that the instrument developed to measure only one construct. The principal component analysis of residuals was used to examine the factors found in the residuals.

As a result of estimation, Table 1 shows that the observed raw variance is 31.9 % and approximates the expected model at 28.9 %. This is in line with Conrad et al. (2011) that stated at least 20 % as consideration for the minimum raw variance dimension. The level of noise measured, or the variance which was not explained in the first contrast shows a 3.5 % value of less than 15 % and is considered to be very good and sufficient (Fischer, 2007; Linacre, 2003). The ratio of the variance described by the measure is 31.9 % with the first component of the variance component at 5.1 % which equals to 6.28: 1, and this exceeds the minimum ratio of 3:1 (Conrad et al., 2011; Embretson & Reise, 2000). The Eigenvalues of 2.9 indicates no significant dimension in the item (Linacre, 2009).

Taken together, these results would seem to suggest that the whole item in the PekkA meets the unidimensionality and construct validity. The results also indicate that the instrument would be able measure vocabulary in the Arabic language effectively.

Table 1: Standardize Residual Variance of PekkA

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

		Empirical			Modeled
Total raw variance in observations	=	83.7	100%		100.0%
Raw variance explained by measure	=	26.7	31.9%		28.9%
Raw variance explained by persons	=	7.0	8.3%		7.6%
Raw variance explained by items	=	19.7	23.5%		21.3%
Raw unexplained variance (total)	=	57.0	68.1%	100%	
Unexplained variance in 1st contrast	=	2.9	3.5%	5.1%	

Fit Statistic

In order to distinguish '*outliers*' or '*misfits*', there are numerous tests used by Rasch analysis. '*Outliers*' or '*misfits*' is generally understood to mean that the estimated value do not meet the whole model fit. In accordance to the Rasch model, the fitness determined by the measures of Outfit Mean Square (MNSQ), the Outfit Z-Standard (ZSTD) and the Point Measure Correlation (PTMEA CORR) indices.

This study uses cut-off values as suggested by Linacre (2003) and Boone et al., (2014) with MNSQ values (0.50<x<1.50), ZSTD (-2.0<x<+2.0) and PTMEA CORR (0.4<x<0.85). If all the criteria are not met with the above indices, it indicates that the items are inappropriate and must be dropped or replaced with other items to ensure that the items measure the respondent's abilities.

The analysis of 57 PekkA's items indicated that 20 items of true and false and 17 items of multiple-choices fall within the acceptable fit range of MNSQ (0.5<y<1.50) (see Table 1). This excludes two items from the sentence completion (F2 and F15) that exceed 1.50 (see Table 2). However, the PTMEA CORR indices were entirely positive. Thus, the ZSTD value can be neglected if the study samples exceed 450 respondents as determined by Linacre (1994). This would appear to indicate that the items were not deleted and accepted. From this investigation, all 57 items fulfilled the item fit criterion and measured what it intends to measure.

Table 2: Misfit Items

Item	Measure	S.E	OUTFIT		PTME
			MNSQ	ZSTD	CORR
YN4	0.65	0.12	1.10	3.06	0.19
YN17	0.40	0.12	0.94	-2.02	0.30
M1	-0.25	0.10	0.92	-2.03	0.38
M7	-0.26	0.10	0.89	-2.64	0.43
F1	-0.35	0.06	1.48	5.32	0.43
F2	-1.84	0.08	1.56	3.52	0.34
F5	-0.61	0.06	0.81	-2.69	0.61
F8	0.08	0.06	1.26	2.58	0.32
F9	0.48	0.07	0.66	-3.19	0.53
F11	0.26	0.07	0.74	-2.72	0.53
F14	0.57	0.07	0.64	-3.26	0.52
F15	0.71	0.07	1.57	3.62	0.33

Table 3: Construct of Misfit Items

Item	Outfit MNSQ	Construct	Item
F2	1.52	Denotative meaning	يصلي المسلمون جَ_____ في المساجد.
F15	1.57		مَدِيْنَةٌ كوالا لومفور عَ_____ةٌ.

Local Independence

In the Rasch analysis, local independence of items requires the items to be independent of each other. This means a correct or wrong response to one item should not lead to a correct

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

or wrong response to another item. The items should only be correlated to the latent traits that the test is measuring as stated by Lord and Norvick (1968).

The analysis shows a standard value of residual correlation at the range of -0.19 to 0.24. This range of value indicates that local freedom requirements complied with the correlation score at less than 0.30 (Balsamo et al., 2014). This finding (see Table 4) implies that all items are independent despite being in the same construct.

Table 4: Local Independence

Correlation	Item Number	Item Number
0.24	F16	F54
-0.25	YN4	F10
-0.24	F2	F12
-0.22	YN17	F10
-0.20	YN7	F10
-0.19	M10	F16
-0.19	M6	F18
-0.19	YN8	F15
-0.19	YN15	F4
-0.19	YN18	F17

Differential Item Functioning (DIF)

DIF exists when an item has a different difficulty level in a group of variables such as gender, age, racial background, economic status, experience, etc. One of the frequently cited literature of construct-irrelevant variance is gender, which has a vital role in differences of vocabulary acquisition (Dörnyei, 2003). Therefore, there is considerable concern about DIF in terms of gender in PekkA's instrument.

From the DIF analysis, there is a difference in response patterns between 143 male students (33.3%) and 340 female students (77.6%). Table 5 below shows that based on gender, there were five items tend to have different responses (items YN2, YN5, YN13, YN18, and F13). These items had the probability score of less than 0.05 and DIF contrast score exceeded 0.64 as suggested by Linacre and Wright (2012).

Table 5: DIF Class Specification by Gender

Item Number	Item Label	Construct	DIF Measure	DIF Contrast	Prob.
2	YN2	Word formation	-1.09	0.67	0.02
5	YN5	Semantic relation	-2.05	0.84	0.03
13	YN13	Semantic relation	-1.49	0.77	0.01
18	YN18	Word formation	0.87	1.37	0.00
50	F13	Denotative meaning	1.42	0.64	0.01

The DIF graph curve (see Figure 1) portrays that the female students found it was easier to answer item YN2 and YN13 (True and False) compared to the male students. Meanwhile, the male students found item YN5 to be easier than the female students. At the same time, the female students found items YN18 and F13 (text completion) to be more difficult than the male students.

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

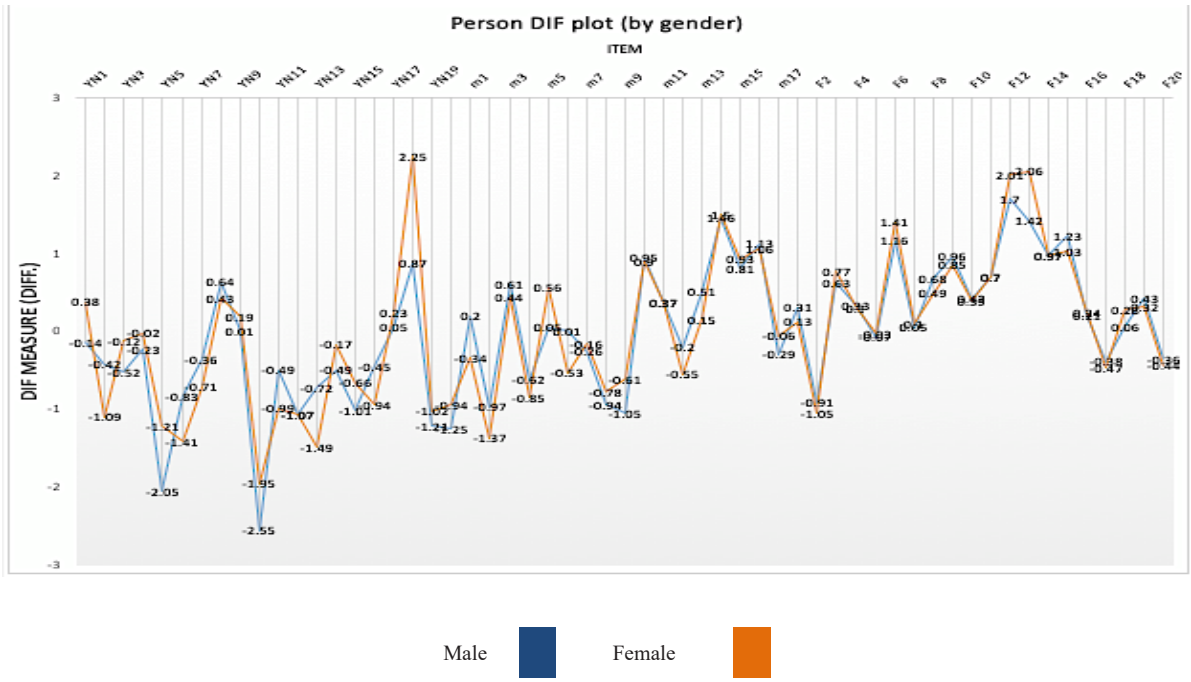

Figure 1: DIF Graph by Gender

However, based on item fit, the scores showed that the whole item was in the acceptable MNSQ, then no item elimination or change was required. From the DIF analysis, it can be concluded males performed better than females. This test was biased slightly in the Yes/No response and text completion.

Person and Item Reliability

The instrument reliability in Rasch measurement model relies on reliability and separation of person and item. Still, the person separation index obtained in order to indicate whether the item is dispersed across the continuum of the logit ability or latent trait. The bigger the person separation index the more likely the respondents will respond correctly to the items. This is useful to determine whether the construct is meaningful and measures what it should measure (Bambang & Wahyu, 2014).

Table 6 below shows the person reliability is 0.81 and the person separation index at 2.06. This index portrayed an acceptable and good value as asserted by Fischer (2007) that the reliability value above 0.94 as excellent, 0.91 to 0.93 as very good and the value of 0.81 to 0.90 is considered good. While the person separation index of PekkA is 2.06 which indicates that the person separation index can be divided into two groups. However, using the strata equation formula (Wright & Master, 2002), the person separation score is 3.08, which showed three groups of persons; excellent, moderate and weak groups.

Meanwhile, item reliability is 0.99 (see Table 6), which is acceptable and considered very strong (Fischer, 2007). The high reliability of item value was due to the wide difficulty range of items and large sample size. This implies that the sample size is sufficient and the respondents responded to the tests consistently (Linacre & Wright, 1994). Table 6 also demonstrates the item separation index is 9.23.

This finding verifies that the items can be grouped into nine segments ranging from too difficult to too easy items. The item separation index shows that the items are excellent and acceptable, based on Fischer (2007) in which a separation index of more than five considered excellent. High item separation index >3 and item reliability >0.9 implied that the person sample is enough to confirm the hierarchy

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

of item difficulty, which is considered as the construct validity of the instrument (Linacre & Wright, 1994).

Collectively, this finding is in line with Bond and Fox (2015); Linacre & Wright (2012) that stated the person reliability index that exceeds than 0.80 along with the item reliability more than 0.90 proves that the sample of the study is sufficient. The evidence presented signifies PekkA as a reliable instrument to be used for a different group of respondents.

Table 6: Person and Item Reliability

	Mean Logit (SD)	Separation	Reliability	Alpha Cronbach
Person	0.49	2.06	0.81	0.81
Item	0.87	9.23	0.99	

Person-Item Map

One of the significant features of the Rasch measurement model is the variable map or person-item map that able to plot the dissemination of persons and items on a single measurement scale. As shown in Figure 1, the maximum level of item measurement was +1.91 logit (SE: 0.15) while the maximum measure of a person was +1.61 logit (SE: 0.30). This range of items described that the scale can be exploited only from +1.61 to -1.26 equal to ±2.87 logit as shown below (Figure 2).

There were items that were too difficult and were left unanswered or person-free at +1.74 (YN18 word formation), +1.83 (F13 denotative meaning) and +1.91 logit (F12 word combination) compared with the most competent person at 1.61 logit. There were also very easy items that were person-free, which showed the item measurement at -1.42 (YN5 semantic relation) and -2.11 (YN10 word combination) logit compared to the least competent test taker at -1.26 logit. This shows that the most difficult items are the text completion tasks (F12 word combination), and the least difficult items were the True and False responses (YN10 word combination). This instrument was found to have a low standard error of measurement and (SE) 0.09 that indicates excellent targeting (Fischer, 2007).

The person-item map also showed that the coded respondent (065PSA) had the highest Arabic vocabulary test ability. The student is a female from the southern region and has earned Grade A in Form Three Assessment for the subject of Arabic language in reference to the respondent code. While the respondent coded (270LTB) was a male student who received a B in Arabic and from the eastern region. The respondent has the lowest ability in this vocabulary test.

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

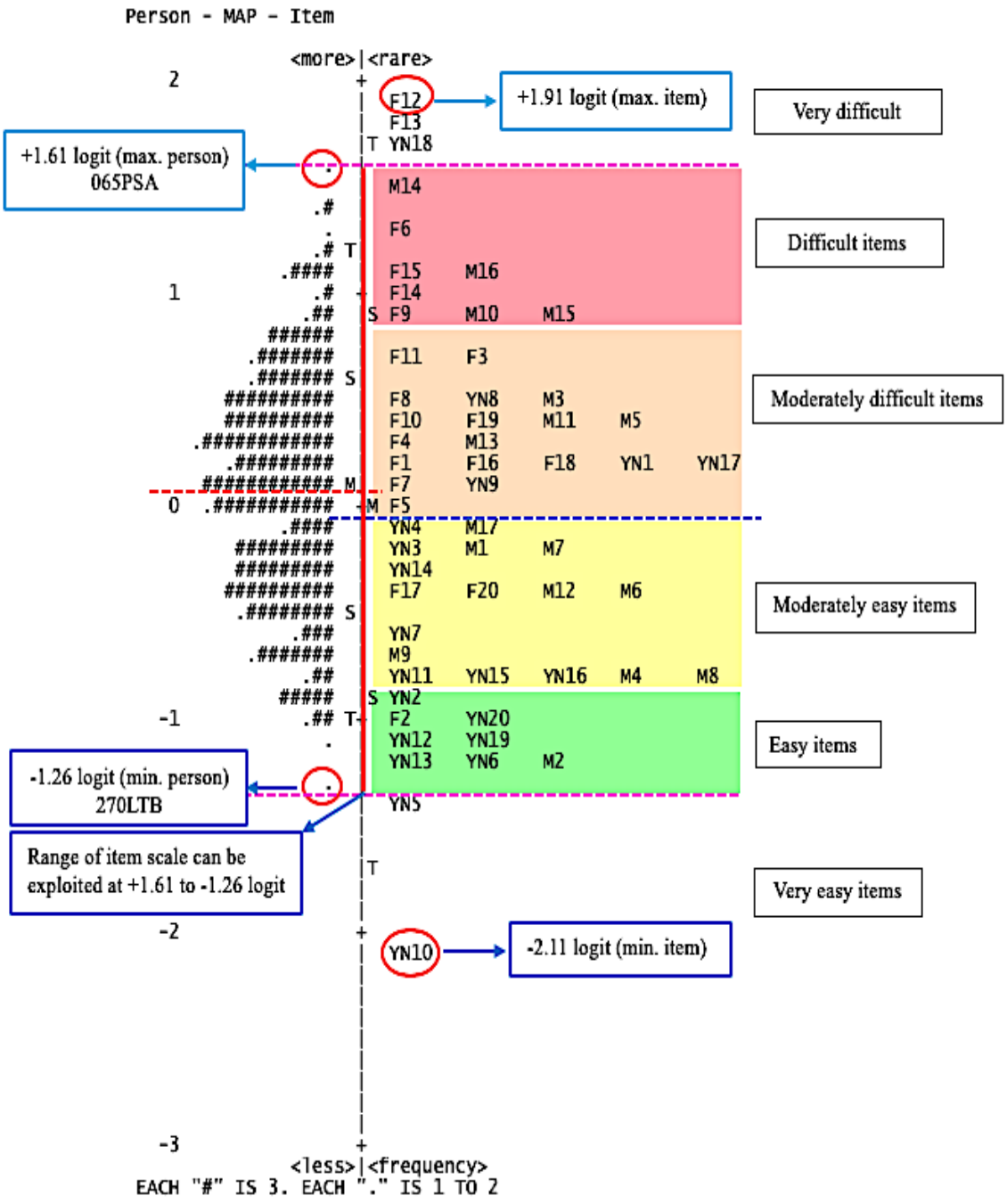

Figure 2: Person-Item Map

CONCLUSION

This study portrayed the robustness of the Rasch measurement model in developing a test instrument for assessing Arabic vocabulary knowledge among Arabic language learners at secondary schools. Through the Rasch analyses, diverse procedures were applied to evaluate the psychometric properties of PekkA which are based on the IRT framework. This study also, uses the latent trait of the Rasch model and postulated an overview of empirical evidence in validity evidence through authentic interval measures that met the requirements of the IRT perspective.

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

Validity evidence as proof of internal consistency was proven through the unidimensionality test. The validity evidence showed the coherence of items and their conformity to the requirement of unidimensionality as stated by Conrad et al., (2011); Linacre (2009); Fischer, (2007); Embretson & Reise, (2000). All 57 PekkA's items were fulfilled the item fitness criterion. This is due to the accepted value of the fit statistic is close to 1. The PTMEA CORR was reported positive and ZSTD value within the range of ±2.0. In term of local independence, all PekkA items are verified independent and being in the same construct at the cut of a point less than 0.30 as suggested by Balsamo et al. (2014).

IRT perspective also implied to examine the presence of DIF for males and females on Pekka's items using the Rasch model. Therefore, the DIF analysis of the study revealed that males performed better than females in responding to the items. There was a slight gender bias found in Yes-No task and text completion, however, no item was eliminated due to the fit statistic evaluation that all items were fit.

Meanwhile, the reliability evidence of the PekkA instrument verified that both person and item measures were acceptable as mentioned by Fischer (2007); Linacre & Wright (1994). The person-item map plotting out the responses of respondents against the items with varying levels of difficulty. This concluded that there was sufficient item and clearly differentiate between the high and low ability respondents.

With the examination of the psychometric properties of the PekkA, it was found that the validity and reliability of the instrument were acceptable to measure Arabic vocabulary knowledge among Islamic secondary school students. It also fulfilled psychometric assumptions in the Rasch measurement model. These imply that PekkA is an empirically tested and validated instrument for evaluating vocabulary knowledge.

Conclusively, the psychometric evidence satisfied the research objectives to examine the item suitability in research instruments that conform to the model and theoretical framework. For sensible usefulness, the difficulty level of the items option should be revised in the attempt to broaden the use of PekkA instrument in other school settings. Together this study provides important insights into Arabic language development specifically in vocabulary research. The development of the PekkA instrument as a diagnostic tool is predicted to be able to recognize the Arabic students' ability in vocabulary knowledge focusing on meaning.

REFERENCES

Abdul Razif Zaini & Mohd Zaki Abd. Rahman. (2015). *Saiz Kosa Kata Bahasa Arab Dalam Kalangan Penuntut Jurusan Pengajian Islam dan Bahasa Arab*. Jurnal Pengajian Islam, 8 (2): 225-238.

Alderson, J. C. (2005). Diagnosing Foreign Language Proficiency: The Interface Between Learning And Assesment. London: Continuum.

Al-Naqah, M. K. (1985). *Ta`alim Al Lughah al`Arabiyyah Linnatiqina Biha*. Saudi Arabia: Universiti Ummul Qura.

Al-Shuwairekh, S. (2001). Vocabulary Learning Strategies Used by AFL (Arabic as a Foreign Language) Learners in Saudi Arabia. Unpublished Ph.D. Thesis, University of Leeds.

Azrilah Abdul Aziz, Mohd Saidfudin Masodi & Azami Zaharim. (2013). *Asas Model Pengukuran Rasch: Pembentukan Skala & Struktur Pengukuran*. Bangi: Penerbit Universiti Kebangsaan Malaysia.

Bachman, L. F. (1990). Fundamental Consideration in Language Testing. Oxford: Oxford University Press.

Balsamo, M., Giampaglia, G. & Saggino,

A. (2014). Building A New Rasch-Based Self-Report Inventory of Depression. Neuropsychiatric Disease and Treatment, 10: 153–165. doi:10.2147/NDT.S53425

Bambang Sumintono & Wahyu Widhiarso. (2014). *Aplikasi Model Rasch Untuk Penelitian Ilmu - Ilmu Sosial. (Edisi Revi.)*. Cimahi, Indonesia: Trim Komunikata Publishing House.

Bond, T. G. & Fox, C. M. (2015). Applying the Rasch Model: Fundamentals Measurement in the Human Sciences (3rd ed.). New York: Routledge.

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). Rasch Analysis in the Human Sciences. Dordrecht: Springer.

Buckwalter, T. & Parkinson, D. (2011). A Frequency Dictionary of Arabic (1st ed.). London: Routledge.

Chapelle, C. A. (1994). Are C-tests Valid Measures for L2 Vocabulary Research?. Second Language Research, 10: 157-187.

Che Radiah Mezah & Norhayuza Mohammad. (2013). *Kosa Kata Arab : Teori Dan Aplikasi*. Serdang: Penerbit Universiti Putra Malaysia.

Conrad, K. J., Conrad, K. M., Dennis, M. L., Riley, B. B. & Funk, R. R. (2011). Validation of Crime and Violence Scale (CVS) to Rasch Measurement Model GAIN Methods Report 1.2. Evaluation Review. Chicago: IL.

Conrad, K. J., Conrad, K. M., Mazza, J., Riley, B. B., Stein, M. A. & Dennis, M. L. (2012). Dimensionality, Hierarchical Structure, Age Generalizability, and Criterion Validity of the GAIN's Behavioral Complexity Scale. Psychological Assessment, 24 (4): 913-924.

Dörnyei, Z. (2003). Attitudes, Orientations, and Motivation in Language Learning: Advances In Theory, Research, and Applications. Language Learning, 53: 3-32. doi:10.1111/1467-9922.53222

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Fischer, W. P. (2007). Rating Scale Instrument Quality Criteria. Rasch Measurement Transaction, 21 (1): 1095.

Gorin, J. S. & Embretson, S. E. (2008). Item Response Theory and Rasch Models. In D. McKay (Ed.), Handbook of Research Methods in Abnormal and Clinical Psychology (pp. 271-292). Thousand Oaks, California: SAGE Publications Inc.

Haastrup, K. & Henriksen, B. (2000). Vocabulary Acquisition: Acquiring Depth Of Knowledge Through Network Building. International Journal of Applied Linguistics, 10: 221-240.

Harun Baharudin, Zawawi Ismail, Adelina Asmawi & Normala Baharuddin. (2014). TAV of Arabic Language Measurement. *Mediterranean Journal of Social Sciences*, 5(20): 2402-2409. doi: 10.5901/mjss.2014.v5n20p2402

Henriksen, B. (1999). Three Dimensions of Vocabulary Development. Studies In Second Language Acquisition, 21: 303-317.

Linacre, J. M. (1994). Sample Size And Item Calibration (Or Person Measure) Stability. Rasch Measurement Transactions, 7 (4): 328.

_____. (2003). Dimensionality: Contrast and Variances. Retrieved June 1st, 2018, from http://www.winstep.com/winman/principalcomponents.html

_____. (2006). A User's Guide to Winsteps Ministeps: Rasch-Model Computer Programs. Chicago, IL: Electronic Publication.

_____. (2009). FACETS Computer Programs Version 3.66.1. Chicago: MESA

ISSN : 2232-0474 | E-ISSN : 2232-0482
www.gjat.my

Press.

Linacre, J. M. & Wright, B. D. (1994). Reasonable Mean Square Fit Values. Rasch Measurement Transactions, 8 (3):370. Retrieved June 1st, 2018, from http://www.rasch.org/rmt/rmt83.html

_____. (2007). A User's Guide to WINSTEPS Ministeps Rasch Model Computer Programs. Chicago: MESA Press.

_____. (2012). A User's Guide to WINSTEPS Ministeps Rasch Model Computer Programs. Chicago: MESA Press.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Malaysia Ministry of Education. (2006). *Al-manhaj al-dirasi lil-lughah al-A'rabiah: al manhaj al-mutakamil lil-madaris al-thanawiyah*. Putrajaya: Bahagian Pendidikan Islamdan Moral.

McCreary, L. L., Conrad, K. M., Conrad, K. J., Scot, C. K., Funk, R. R. & Dennis, M. L. (2013). Using the Rasch Measurement Model in Psychometric Analysis of the Family Effectiveness Measure. Nursing Research, 62(3): 149-159. doi:10.1097/NNR.0b013e31828eafe6

Milton, J. (2009). Measuring Second Language Vocabulary Acquisition. Bristol: Multilingual Matters.

Nation, I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Qian, D. D. (2002). Investigating The Relationship Between Vocabulary Knowledge and Academic Reading Performances: An Assessment Perspective. Language Learning, 52(3): 513-536. doi:10.1111/1467-9922.00193

Qian, D. D. & Schedl, M. (2004). Evaluation of an In-Depth Vocabulary Knowledge Measure for Assessing Reading Performance. Language Testing, 21 (1): 28-52.

Schmitt, N. (2000). Vocabulary in Language Teaching. Cambridge: Cambridge University Press.

Siti Rahayah Ariffin. (2008). *Inovasi Dalam Pengukuran dan Penilaian Pendidikan.* Bangi: Fakulti Pendidikan, Universiti Kebangsaan Malaysia.

Smith, E.V., Conrad, K. M., Chang, K. & Piazza, J. (2002). An Introduction to Rasch Measurement for Scale Development and Person Assessment. Journal of Nursing Measurement, 10 (3): 189-206.

Tha`imah, R.A. 1989. *Taalim al arabian li ghairi an natiqin biha: Manahijuhu wa asalibuhu*. Rabat: Mansyurat ISESCO.

Wesche, M. & T. S. Paribakht. (1996). Assessing Second Language Vocabulary Knowledge: Depth vs Breadth. Canadian Modern Language Review, 53: 13-39.

Wright, B. D. & Masters, G. N. (2002). Number of person or item strata. Rasch Measurement Transaction, 66: 888.

ZM Maskor, H Baharudin, MA Lubis, NK Yusuf. (2016). Teaching and Learning Arabic Vocabulary: From a Teacher's Experiences, Creative Education, 7(03):482-490. doi: 10.4236/ce.2016.73049

ZM Maskor, H Baharudin, MA Lubis. (2018). Measurement Validity And Reliability Of The Productive Vocabulary Knowledge Instrument For Arabic Learners In Malaysian Secondary Schools, Advanced Science Letter, 24(5): 3423-3426. doi: https://doi.org/10.1166/asl.2018.11-400