THE EFFECTS OF TEST CHARACTERISTICS ON THE HIERARCHICAL ORDER OF READING SKILLS

[1]Kamal J I Badrasawi, [2]Noor Lide Abu Kassim & [3]Nuraihan Mat Daud

[1] Department of Curriculum & Instruction, Kulliyah of Education, International Islamic University Malaysia, Malaysia
[2] Department of Community Nursing & Health Care for Mass Gathering, Umm al-Qura University, Mecca, Saudi Arabia & Kulliyyah of Dentistry, International Islamic University Malaysia and Department of Language & Literacy, Kulliyah of Education, International Islamic University Malaysia
[3]Kulliyyah of Languages and Management International Islamic University Malaysia, Malaysia

[2]Corresponding author: dr.noorlide@gmail.com

ABSTRACT

Purpose – The study sought to determine the hierarchical nature of reading skills. Whether reading is a 'unitary' or 'multi-divisible' skill is still a contentious issue. So is the hierarchical order of reading skills. Determining the hierarchy of reading skills is challenging as item difficulty is greatly influenced by factors related to test characteristics. To examine the interaction between these factors and item difficulty, and determine the possibility of such a hierarchy, this study used the multifaceted Rasch approach.

Methodology – In this descriptive study, a 42-MCQ reading test was administered to 944 ESL lower secondary students, randomly selected from eleven Malaysian national-type schools in the Federal Territory of Kuala Lumpur and the state of Selangor. These student populations were selected as the development of reading ability was considered critical at this stage of schooling. The reading test items were identified according to the following aspects: Reading Skill Areas (*Interpreting information, making Inference, Understanding figurative language, Drawing conclusions, Scanning for details and*

Finding word meanings), Context Type (*Linear and Non-Linear*), and Text Type (*Ads, Notice, Chart, Story extract, Short message, Poem, Short news report, Brochure, Formal letter, Conversation, Long passage and weather forecast*). Applying the Many-Facet Rasch model of measurement, the study analyzed student responses to the test items with the help of FACETS, version 3.7.1.4.

Findings – The findings showed that context types, skill areas, and text types differed in difficulty ($p<.01$), with those items that required understanding and interpretation being more demanding. Test items based on linear contexts were more difficult than those based on non-linear contexts. *Understanding figurative language* was found to be the most difficult skill followed by *Making inference* and *Interpreting information.* The easiest reading skill was *Scanning for details,* followed by *Finding word meanings.* The reading skill, *Drawing conclusions,* was close to the average difficulty level. The findings also indicated that texts that were longer and had more information tended to be more difficult.

Significance – This study has also shed new light on the theory and practice of reading. The findings support the hierarchical nature of reading skills. Different reading skills were found to exert differential cognitive demands, and those which required higher cognitive ability were more difficult for learners to acquire and perform. Understanding the hierarchy of reading skills will help language teachers to target their teaching more effectively; course designers to produce more appropriate teaching and learning materials; and test writers to develop test items that better meet students' reading competencies.

Keywords: Reading hierarchy; reading skills; many-facet Rasch analysis; FACETS.

INTRODUCTION

Given the importance of reading skills in language development, a considerable amount of research has tried to identify the nature of reading skills (Alderson, 2005; Grabe & Stoller, 2002; Hedgcock & Ferris, 2009); how they develop: how they should be taught; and how they should be assessed (Alderson, 2005; McNamara, 1996).

Yet, defining them is difficult and varied (Grabe & Stoller, 2002) as there is no consensus on the nature of these language skills and how they are developed (Alderson, 2005; Hedgcock & Ferris, 2009; Sainsbury, Harrison, & Watts, 2006; Urquhart & Weir, 1998). To date, two main opposing views on reading skills dominate: reading as an 'indivisible' or a 'unitary' skill, and reading as a 'multi-divisible' skill (Alderson, 2005; Weir & Porter, 1994).

The former view implies that reading skills do not have clear separable and identifiable sub-skills or components (Alderson & Lukmani, 1989; Alderson 1990 a, 1990 b; Bachman, 1990; Rost, 1993; Weir & Porter, 1994). However, Weir and Porter (1994) argued that a bi-divisible view of reading is plausible as 'vocabulary' seems to be a separate component from reading comprehension as evidenced in a number of quantitative researches. Weir, Huizhong and Yan (2002) quoted several studies in which these two components of reading were identified (Berkoff, 1979; Carver, 1992; Farr, 1968; Guthrie & Kirsch, 1987). Alderson (2005) also highlighted that the common view in the research literature was that reading could be seen as comprising the following two components: decoding (word recognition) and comprehension. Hughes (2003) explicated that despite the issue of the existence of subskills in reading, the reading test must include samples of skills that are relevant to the test purpose. Hence, it is important that reading assessment should be guided by a clear reading theory that defines the reading skills accurately. This helps the measurement and interpretation of students' performance in the specific reading skill of concern much more precisely (Engelhard, 2001).

The latter view maintains that a particular reading skill has separable and identifiable sub-skills. It could be divided into different subskills as greatly evidenced in the literature (See Davis, 1968; Farhadi & Moeini, 2005; Farhady & Hessamy, 2005; Kim, 2009; Matthews, 1990; Munby, 1978; Sainsbury, Harrison, & Watts, 2006; Spearritt, 1972; Weir, Hughes, & Porter, 1990). In this respect, the distinction between high and low order skills of reading and the relationship between them is essential for a better understanding of the nature of a reading skill, for constructing valid items to test reading ability as well as for planning syllabuses (Alderson, 2005; Lumley, 1993; Weir, Hughes, & Porter, 1990).

One important issue that clouds the multi-divisible view of reading skills is the controversy surrounding the number of skills and the hierarchical ordering of these skills or sub-skills. Questions like "*what such skills might consist of and how they might be classified, acquired, taught and tested*" (Alderson, 2005, p.10) need to be answered. Hudson (2007) argued that both L1 and L2 research consistently failed to support a "strictly hierarchically ordered reading skills" (p. 103) position. There is no clear evidence that reading could be divided into high and low order skills (Alderson 1990a, 1990b; Hudson, 2007; Rost, 1993). Despite the different views about the nature of reading, the notion of skills and sub-skills is influential (Alderson, 2005). This is seen through the use of various taxonomies which are used in teaching reading as well as for testing it (See Alderson & Lukmani, 1989; Grabe, 1991; Hudson, 2007; Matthews, 1990; Pearson & Johnson, 1978; Urquhart & Weir, 1998; Vacca & Vacca, 2008).

Despite their seeming utility, the use of these taxonomies is not without criticisms (Alderson, 2005). For instance, Matthews (1990) argued that Munby's taxonomy (1978) is a knowledge-based taxonomy rather than a skill-based one, and thus the latter taxonomy should be disregarded. Moreover, some taxonomies, such as Munby's, Barrett's and Bloom's were developed largely based on theoretical assumptions and not on empirical frameworks (Hudson, 2007). It has also been pointed out that "skills hierarchies should not be interpreted in an *a priori* fashion, as the field has not reached consensus on what constitutes higher or lower order skills, which are relative and subject to the influence of the context for reading" (Hedgcock & Ferris, 2009, p.38).

In the testing of reading ability, it was found that item difficulty was influenced by a number of factors other than the inherent difficulty of the skill (Alderson, 2005; Bachman, 1990; Day & Park, 2005; Kobayashi, 2005; McNamara, 1996). These included factors such as question type, context type, question format, cognitive demand, explicitness and implicitness of information, students' test-taking skills (Alderson, 2005; McKenna & Stahl, 2009; Pearson & Johnson, 1978), text type as well as text length (Scheuneman & Gerritz, 1990). With regard to the type of question format, a multiple choice question is influenced by its stem length, stem content words, structure of

options, length of correct answer and distractors (Alderson, 2005). Furthermore, Pearson and Johnson (1978) found that question types varied in difficulty based on the explicitness and implicitness of question information. The research literature also showed that item characteristics and the interaction of these characteristics affected the item difficulty (Alderson, 2005). Alderson further elaborated that the factors or influences that affected item difficulty or made the task more demanding should be controlled; otherwise, they could be a risk to test validity. Hence, for accurate and valid conclusions of skills hierarchies it is essential to account for these factors and model their influence on item difficulty.

One important concern in investigating skills hierarchies and item difficulty is in the quantitative or qualitative method used. Daftarifard and Lange (2009) noted that judgmental analysis of item difficulty was insufficient, as many studies had found discrepancies between the hypothesized order based on expert judgment and the item difficulty estimates from empirical analysis (see also Weir et al., 2002). Such a judgment should be empirically tested (Lumley, 1993). The estimation of item difficulty without modeling the effects of the factors mentioned earlier is equally problematic. Hence, a multifaceted approach using the Many-facet Rasch analysis has been recommended to examine item difficulty. Such an approach would consider the influence of other related variables on item difficulty (Daftarifard & Lange, 2009). According to them,

> …given this lack of correspondence, we propose that notions of items complexity require careful distinctions between the qualitative and quantitative aspects of reading theory. For instance, it may be necessary to distinguish between the complexity of a concept and the complexity of the question designed to assess this concept. Rasch [analysis] is likely to remain the tool of choice in this research, but it seems likely that multi-facetted approaches will be needed to accommodate both types of complexity simultaneously (p. 1212)

Given the possibility of examining the *"complexity of a concept and the complexity of the question designed to assess [the] concept"* using the Many-facet Rasch analysis, this study explored

the hierarchical assumption of reading skills for support of a developmental hierarchy of reading ability using the Many-facet Rasch model (Linacre, 1989, 2014a). It also examined the influence of particular item characteristics on item difficulty. The findings of this study would be able to shed light on the issue of the hierarchy of reading skills, and the robustness of the Many-facet Rasch model in this regard.

METHODOLOGY

Participants and Sampling

This study employed the descriptive design method (Fraenkel, Wallen, & Hyun, 2012; Gay, Mills, & Airasian, 2000; Keeves, 2004). Here, a dual-purpose instrument comprising two sections was used; Section A asked some questions about students' demographic information and Section B comprised a 42-MCQ item reading test.

The participants consisted of lower secondary ESL students (i.e.; Forms 1, 2 and 3; 13 to 15 year olds) in Malaysia. The lower secondary level begins at the end of the primary level which lasts six years and before the upper secondary level which lasts two years. This population was selected as the development of reading ability was considered critical at this stage of learning. It was also considered one of the foremost components and indicators of being literate (McGee & Richgels, 2004); it was also seen as helping students to succeed in their studies (Holme, 2004), and to perform their daily and personal affairs more effectively (Vacca & Vacca, 2008).

A representative sample was chosen from 11 national-type secondary schools which were randomly selected from the Federal Territory of Kuala Lumpur and Selangor state in Malaysia. For each school, 30 students were again randomly selected from each Form (grade level) giving a total of 990 students (30 students x 3 Forms per school x 11 schools). However, the total number of students included in the final analysis was 944 out of the 990 who were selected. When the test papers were examined to ensure the integrity of the data collected, some test papers were found blank, so they were excluded; and some

of the selected students were not available on the days the data were collected.

Instrument

A 42-MCQ English reading test was used in the study. Ingebo (1997) recommended the use of 40 items as an acceptable number for tests using Rasch analysis. The instrument for this study was developed as follows. First, three sets of past English language reading tests (for the years 2004, 2005, and 2007) used to assess reading comprehension at the national level (*Penilaian Menengah Rendah* [PMR]) were selected. The PMR is a national standardized examination conducted at the end of the lower secondary education for Form 3 students. The papers also included specific tasks and competencies associated with English reading literacy that students were expected to possess over time in the lower secondary period. The reading skills included in the PMR exam papers were therefore, those skills that had been taught to students in Forms 1, 2, and 3, given the spiral curriculum design adopted by the Malaysian Ministry of Education.

The three PMR tests were similar in test format and number of items. Each paper included 60 multiple-choice items (40 for comprehension questions and 20 for grammar) with four options. Based on the syllabuses of English for Forms 1-3, the researchers, with the help of experts in the English language comprising two university lecturers and two school teachers of English, analyzed the three sets of PMR tests and came out with item content descriptors. In doing so, they were able to pinpoint the level of item difficulty, skill/sub skills and grade level the test items represented. It is worthwhile to mention that these PMR tests were developed by content experts and teachers from the field who were appointed by the Ministry of Education; therefore, content validity was not considered an issue.

Data Analysis

To select the most appropriate test, the MFORMS for concurrent analysis (Linacre, 2014a) was used to link the three tests. To allow for common item linking, a set of 20 grammar items were included in all the three tests. Each set was administered to different groups

of school children (*n*=269) selected from Forms 1, 2, and 3 in three national type secondary schools in Kuala Lumpur, Malaysia. From the concurrent analysis, the 2007 exam paper was found to be appropriate to be used in the final study because it had items that fit the Rasch Model, measuring different reading skills with different difficulty levels. However, language experts, one from the English Department and another from the Faculty of Education at the International Islamic University Malaysia, suggested carrying out certain modifications on several of the items and adding others from the other two exam sets for sufficient coverage of the various skills being investigated in the main study. The final test included 60 multiple-choice items with 4-options (42 for comprehension questions and 18 for grammar). In the analysis, only the 42 reading comprehension items (see Table 1) were analyzed using the Many-facet Rasch analysis.

Table 1

Test Items: Reading Subskills, Context types, and Text Types

Variables	Number of Items (n=42)
Skill	
Interpreting information	5
Making inference	6
Understanding figurative language	2
Drawing conclusions	8
Scanning for details	6
Finding out word meanings	15
Context type	
Linear	30
Non-linear	12
Text type	
Ads	2
Notice	3
Chart	2
Story extract	3
Short message	1
Poem	4

(continued)

Variables	Number of Items (n=42)
Short news report	2
Brochure	5
Formal letter	6
Conversation	8
Long passage	5
Weather forecast	1

These test items represented the following skill areas: *Interpreting information* (5 items), *Making inference* (6 items), *Understanding figurative language* (2 items), *Drawing conclusions* (8 items), *Scanning for details* (6 items), and *Finding out word meanings* (15 items). With regard to context type, 30 linear items and 12 non-linear items were identified. In terms of text type, the distribution of items was as follows: *Ads* (2 items), *Notice* (3 items), *Chart* (2 items), *Story extract* (3 items), *Short message* (1 item), *Poem* (4 items), *Short news report* (2 items), *Brochure* (5 items), *Formal letter* (6 items), *Conversation* (8 items), *Long passage* (5 items), and *Weather forecast* (1 item). The items or tasks vary in terms of difficulty and were targeted to assess the reading skills of students with different levels of ability.

The Many-facet Rasch Model (Linacre, 1989, 2014a), an extension of the basic Rasch Model, was used for analysis as it could allow for other facets other than person ability and item difficulty to be modelled and evaluated (See Linacre, Engelhard, Tatum, & Myford, 1994; Lunz & Wright, 1997; McNamara, 1996). And since item characteristics exerted an influence on item difficulty (Alderson, 2005), the multifaceted measurement approach was considered appropriate. For this study, the computer program, FACETS version 3.7.1.4 (Linacre, 2014 b) was used. The analysis was conducted in two FACETS runs to determine the descriptive summaries of effects of item and test characteristics (Linacre, 2014b). In the first run, person ability estimates and item difficulty estimates were calibrated.

The subset connection which showed the link between all elements in the analysis (persons, items, text type and context type) indicated that all elements of analysis were estimated in an unambiguous frame of reference. The mean standardized residuals (0.02) and the sample standard deviation. (1.01), indicated that the data fit the Rasch

Model. The Rasch person reliability coefficient was acceptable (.88) and close to .90, which is expected in high stakes national level testing. All items showed Infit Mean square (MNSQ) values between 0.70 and 1.30, thus showing good fit and no unexpected randomness in responses. It is important to highlight that fit statistics help determine the quality of the collected data and suitability of the items used in the measure (i.e., the test in this case) (Bond & Fox, 2015). The Mean-square fit statistics would show the size of "the randomness, i.e., the amount of distortion of the measurement system" (Linacre, 2002, P.878). In the second run, the items were anchored at the item difficulty measures derived in the first run. Item characteristics that were expected to influence item difficulty (i.e., the context type, text type, and skill areas associated with the items) were then estimated and evaluated in this second run.

RESULTS

The first Facet run showed that the measures of item difficulty estimates spanned about four logits (-1.98 to +2.07 logits). The point measure correlation (PTMEA CORR.) coefficients (which is similar to the point-biserial correlation) for the 42 items were positive and almost all were above 0.3, indicating that the items were effectively discriminating between persons with high ability and those with low ability (Bond & Fox, 2015). For the Infit MNSQ, all items were within the recommended range (0.7-1.3), implying that all the items were productive and meaningful for measurement (Bond & Fox, 2015; Wright, Linacre, Gustafson, & Martin-Lof, 1994). The mean of the Infit MNSQ was 0.99 logit, close to the expected value of 1.00, and the standard deviation was very small (0.11 logit). The high reliability of item difficulty measures (.99) indicated that the ordering of item difficulty estimates was highly replicable with other comparable sample of students and that the items were well-separated in terms of difficulty. The item separation index was 12.49, indicating that the items could be divided into at least 13 difficulty levels. The analysis also showed that the mean for person ability was 0.48 logits and standard deviation was 1.15 logits.

The measures of person ability spanned about six logits (+4.22 to -2.34 logits). For the person fit statistics, the Infit MNSQ indicated that only 38 (4.02%) persons were under fit i.e., with values above

1.3; however, most values were not far departed from 1.3. The reliability of person ability measure was also high at .88, suggesting that it was highly likely that the ordering of students could be replicated with similar items of the same difficulty. The person separation index was 2.66, indicating that the reading test could divide the students into three levels of ability. Figure 1 shows the distribution of students (i.e., the stars in column 2) and the items (i.e., the numbers in column 3) on the same interval scale. The upper part of the scale indicates the most able students who answered most of the items correctly, while the lower part shows the least able students, with more incorrect answers. Items most often correctly answered are positioned towards the lower part of the scale and the least correctly answered ones are positioned towards the upper part of the scale.

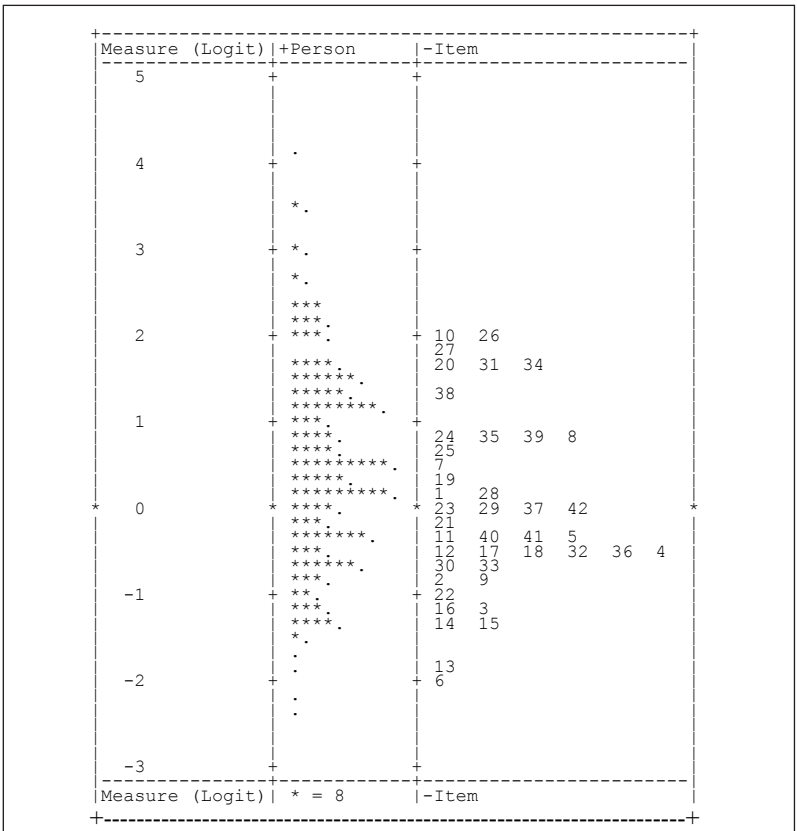

Figure 1. Distribution of Items and Persons on the Logit Scale.

In the second run, the difficulty estimates for context type (linear or non-linear); skill area (*Interpreting information, Making inference, Understanding figurative language, Drawing conclusions, Scanning for details and Finding out word meanings*); and text type (*Ads, Notice, Chart, Long passage, Weather forecast, Conversation, Newspaper report, Story extract, Short message, Poem, Brochure, and Formal letter*) were calibrated, with item difficulty values anchored to the ones derived from the first run. Facets 2 (context) and 3 (skill area) were centred (i.e., mean=0.0) while the fourth facet (text type) was non-centred in this analysis. Figure 2 gives a graphic summary of item difficulty and examinee ability distribution, location of reading skill categories, and context types as well as text types. The results of the analysis showed that item context types were not equally difficult. Linear texts tended to be more difficult (0.38 logit) than non-linear texts (-0.38 logit), a difference of about one logit.

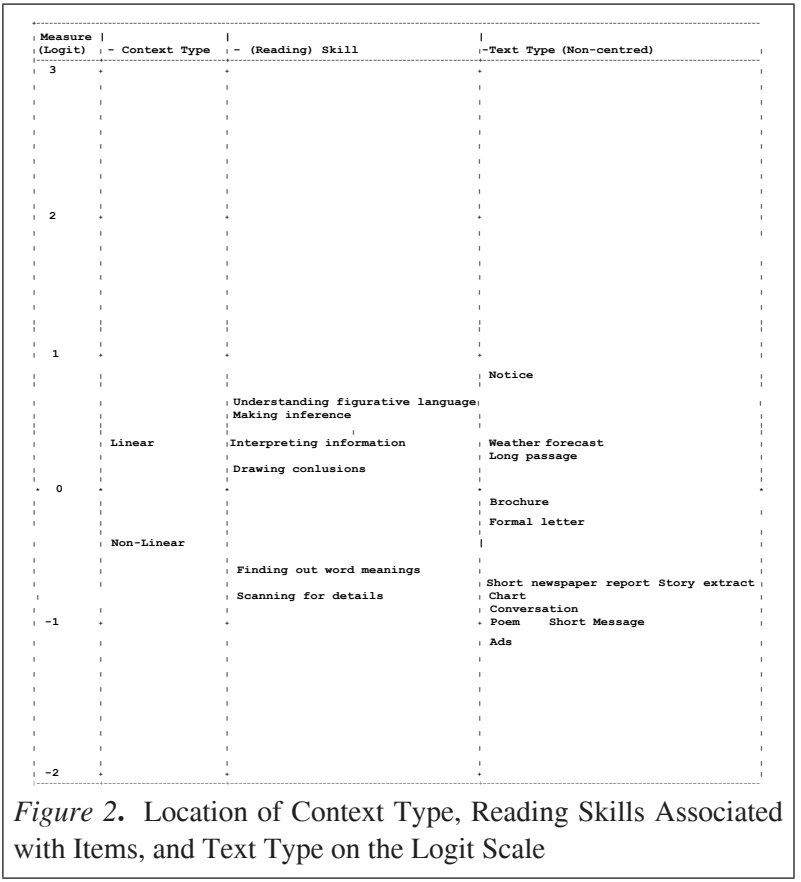

Figure 2. Location of Context Type, Reading Skills Associated with Items, and Text Type on the Logit Scale

Figure 2 reveals that skill categories too did not have the same difficulty level. The skill categories were ordered from the most difficult at the top and the least difficult at the bottom of the scale. Four of the skills were located above the item mean (0.0 logit) and the other two below it. The skill measures showed that the most difficult skill is *Understanding figurative language* (0.56 logit), followed by *Making inference* (0.41 logit), and *Interpreting information* (0.37 logit). The easiest skill was *Scanning for details* (-0.81 logit) and was followed by *Finding out word meanings* (-0.60 logit). On the other hand, *Drawing conclusions* (0.07 logit) was close to the average difficulty level. With a reliability index of 1.00 and a chi-square 2082.3 with 5 df, significant at $p < .01$, it could be concluded that the skill categories were not equally difficult. In terms of fit statistics, all the skills fit the expectations of the Rasch model as they fell within the recommended range of Infit MNSQ of 0.7 to 1.3 (see Table 2) (Bond & Fox, 2015).

With regard to text type, longer texts that contained more complex sentence structures and information (such as long passages, formal letters, brochures) tended to be more difficult than shorter and less cognitive demanding texts, such as advertisements, charts, and short messages.

Table 2

Reading Skill Areas Measurement Report

Reading Skill Area	Observed Score (Count)	Observed Average (Fair-M Average)	Measure (S.E.)	Infit MNSQ	Outfit MNSQ
Understanding Figurative Language	938 (1858)	0.50 (0.46)	0.56 (0.05)	1.00	1.00
Making Inference	3018 (5574)	0.54 (0.50)	0.41 (0.03)	0.96	0.94
Interpreting Information	2398 (5574)	0.43 (0.51)	0.37 (0.03)	1.02	1.03
Drawing Conclusions	4116 (7432)	0.55 (0.59)	0.07 (0.02)	0.99	0.99
Finding out Word Meanings	9425 (13935)	0.68 (0.73)	-0.60 (0.02)	1.01	1.02
Scanning for Details	2947 (4645)	0.63 (0.77)	-0.81 (0.03)	1.00	1.00

DISCUSSION

The results obtained from the Many-facets analysis showed the order of item difficulty measures with regard to skill areas as well as item context types. The most demanding reading skill areas were *Understanding figurative language,* followed by *Making inference*, and *Interpreting information*. The easiest was *Scanning for details* and followed by *Finding out word meanings* and *Drawing conclusions.* Rubin (1993), pointed out that questions that required understanding and interpretation would be difficult for children because readers had to possess "problem-solving ability and be able to work at various levels of abstraction" (p. 196). He added that, to some extent, skill of interpretation depended on the students' ability in skill of inference. For the easy skills, they may represent the literal level, the easiest level of reading comprehension in Barret's taxonomy (Day & Park, 2005; Dupuis, Lee, Badiali, & Askov, 1989; Pearson & Johnson, 1978). This level was described as "an understanding of the straightforward meaning of the text, such as facts, vocabulary, dates, times, and locations.

Questions of literal comprehension can be answered directly and explicitly from the text" (Day & Park, 2005, p. 3). Alderson (2005), for example, also maintained that questions might vary from easy to difficult as a result of cognitive demand; questions which required searching for specific facts were usually less difficult than questions that required synthesis, analysis, or inference. In this respect, Pearson and Johnson (1978) categorized question type into three levels ranging from easy to most difficult: textually explicit, textually implicit and script based. They highlighted that textually explicit questions were "those where both the question information and the correct answer are found in the same sentence. Textually implicit questions, on the other hand, require respondents to combine information across sentences. Script-based questions require readers to integrate text information with their background knowledge since correct responses to the questions cannot be found in the text itself" (p. 87).

The findings on the hierarchy of reading subskills in this study were consistent with those of Hessamy (2013) who concluded that there was the possibility of getting "empirically-based hierarchies of difficulty and importance among the subskills." In other words, a

hierarchy of difficulty of subskills as higher order and a lower order could exist. Hessamy (2013) ordered the examined reading subskills from the easiest to the most difficult as: identifying writer's views/ claims, understanding specific information, identifying main idea, and extracting information from a text to put into diagrammatic representation. This finding supported the view that reading skills could not be a unitary skill.

Additionally, in the current study, the results showed that linear contexts were more difficult than non-linear contexts. Linear contexts were more difficult possibly because most texts of this type were long (four to five paragraphs) and more cognitively demanding, i.e., most of the information was not explicitly stated, and they required students to spend much time on understanding the texts and applying appropriate skills to answer the questions. Whereas, the non-linear contexts included short texts which required less time to read and to figure out the answers. Most items (67%) of this type could be answered easily because the information was explicitly stated in the texts. Of course, not all items of linear contexts were difficult, and not all items of non-linear contexts were easy.

The additional empirical evidence that reading skill had multi-divisible sub-skills with different difficulty levels could help and guide language teachers, course designers, and test item developers to better teach and produce teaching materials as well as test items that would be able to meet their students' reading competencies (Hessamy, 2013). Hughes (2003) also maintained that despite the issue of the involvement of subskills in reading, the reading test would have to include samples of the subskills relevant to the test purpose. Brown (2003) further expounded that the skills used in reading were essential considerations in the assessment of reading ability. It is worthwhile to add that most reading models refer to reading skills or sub processes for profiling purposes, and so language learners should be tested on a range of relevant skills or strategies (Alderson, 2005).

CONCLUSION

The results of the study support the notion that different reading skills exert differential cognitive demands. Those that require higher order thinking skills, such as making inference and interpreting

information, are more difficult than those requiring lower order skills, such as finding out word meanings and scanning for information. The results of the Many-facet anlysis provide the much needed evidence that there is a strong possibility of a hierarchy of reading skills. In this respect, factors that affect item difficulty or order should be estimated to get a more accurate picture of their influence. It will be beneficial to analyze test items and examine features that influence item diffculty before they are administerd to students. Future research that examines the interaction between different reading skills and their effect on language learning and test performance will also greatly benefit reading theory and pratice.

ACKNOWLEDGMENTS

We would like to thank the Malaysian Ministry of Education and the schools that participated in this study for their cooperation and support. This study was funded under the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education, Malaysia: [IIUM/504/RES/G/14/3/05/FRGS 0207-70].

REFERENCES

Alderson, J. (2005). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*(2), 253-270.

Alderson, J. (1990 a). Testing reading comprehension skills (part one). *Reading in a Foreign Language, 6*(2), 425-438.

Alderson, J. (1990 b). Testing reading comprehension skills (part two): Getting students to talk about taking a reading test. *Reading in a Foreign Language, 7*(1), 465-503.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Berkoff, N. A. (1979). Reading skills in extended discourse in English as a Foreign Language. *Journal of Research in Reading, 2*(2), 95-107. https://doi.org/10.1111/ j.1467-9817.1979.tb00197.x.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.

Brown, D. (2003). *Language assessment: Principles and classroom practices*. United States of America: Longman.

Carver, R. P. (1992). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate? *Reading Research Quarterly*, 347-359. https://doi.org/10.2307/747674

Daftarifard, P., & Lange, R. (2009). Theoretical complexity vs. Rasch item difficulty in reading tests. *Rasch Measurement Transactions, 23*(2), 1212-1213.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, 499-545.

Day, R. R., & Park, J. (2005). Developing reading comprehension questions. *Reading in a Foreign Language, 17*(1), 60-73.

Dupuis, M., Lee, J., Badiali, B. J., & Askov, E. (1989). *Teaching reading and writing in the content areas*. Scott, Foresman.

Engelhard, G. (2001). Historical view of influences of measurement and reading theories on the assessment of reading. *Journal of Applied Measurement*, 2(1), 1-26.

Farhadi, H., & Moeini. H. R. (2005). Construct validation of reading comprehension skills. *Quarterly Journal of Humanities*. Al-Zahra Univesity, 25-54.

Farhady, H., & Hessamy, G. R. (2005). Construct validity of L2 reading comprehension skills. *Iranian Journal of Applied Linguistics, 8*(2), 29-53.

Farr, R. (1968). *The convergent and discriminant validity of several upper level reading tests.* A paper presented at the *Yearbook of the National Reading Conference*, 17, 181-191

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2012). *How to design and evaluate research in education*. New York: McGraw-Hill.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2000). *Educational research: Competencies for analysis and application* (6th ed.). Upper Saddle River, NJ: Merrill & Prentice-Hall.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*(3), 375-406. https://doi.org/10.2307/3586977.

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading*. New York: Longman.

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology, 79*(3), 220. https://doi.org/10.1037/0022-0663.79.3.220.

Hedgcock, J., & Ferris, D. R. (2009). *Teaching readers of English: Students, texts, and contexts*. UK: Routledge.

Hessamy, G., & Sadeghi, S. (2013). The relative difficulty and significance of reading skills. *International Journal of English Language Education*, *1*(3), 208-222._http://dx.doi.org/10.5296/ijele.v1i3.4017

Holme, R. (2004). *Literacy: an introduction* .UK: Edinburgh University Press Ltd.

Hudson. T. (2007). *Teaching second language reading*. Oxford: Oxford University Press Oxford.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Ingebo, G. (1997). Probability in the measurement of achievement: Chicago: MESA Press.

Keeves, J. P. (2004). *Educational research, methodology, and measurement: An international handbook* (2nd ed.). England: Pergamon Press.

Kim, A. (2009). Investigating second language reading components: Reading for different types of meaning. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 9 (2,)*, 1-27.

Kobayashi, M. (2005). *An investigation of method effects on reading comprehension test performance.* A paper presented at the Proceedings of the 4th Annual JALT Pan-SIG Conference, Tokyo.

Linacre, M. J. (1989). *Many-facet Rasch measurement*. Chicago: Mesa Press.

Linacre, M. J. (2002). What do infit and outfit, Mean-square and standardized mean? *Rasch Measurement Transactions*, *16* (2), 878.

Linacre, M. J. (2014a). A user's guide to FACETS [Rasch-Model Computer programs program manual 3.71. 4]. Chicago: MESA Press.

Linacre, M. J. (2014 b). Facets for Windows (Version 3.71.4) [Computer Software and manual]. Chicago: MESA Press.

Linacre, M., Engelhard, G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*(6), 569-577. https://doi.org/10.1016/0883-0355 (94)90011-6.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing, 10*(3), 211-234. https://doi.org/10.1177/026553229301000302.

Lunz, M. E., & Wright, B. D. (1997). Latent trait models for performance examinations. *Applications of latent trait and latent class models in social sciences.* Edited by J. Rost and R. Langeheine (pp. 80-88). Münster, New York, München, Berlin: Waxmann, 80-88.

Matthews, M. (1990). Skill taxonomies and problems for the testing of reading. *Reading in a Foreign Language, 7*(1), 511-517.

McGee, L. M., & Richgels, D. J. (2000). *Literacy beginnings: Supporting young readers and writers*. Boston: Allyn and Bacon.

McKenna, M. C., & Stahl, K. (2009). *Assessment for reading instruction*. New York: Guilford Press.

McNamara. T. F. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Pearson, P. D., & Johnson, D. D. (1978). *Teaching reading comprehension*. New York: NJ Holt, Rinehart and Winston.

Rost, D. H. (1993). Assessing different components of reading comprehension: Fact or fiction? *Language Testing, 10*(1), 79-92. https://doi.org/10.1177/026553229 30100 0105.

Rubin, D. (1993). *A practical approach to teaching reading*. London: Allyn & Bacon.

Sainsbury, M., Harrison, C., & Watts, A. (2006). *Assessing reading: From theories to classrooms*. UK: Cambridge.

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131. https://doi.org/10.1111/j.1745-3984.1990.tb00737.x.

Spearritt, D. (1972). Identification of sub-skills of reading comprehension by maximum likelihood factor analysis. *ETS Research Bulletin Series, 1972*(1), 1-24. https://doi.org/10.1002/j.2333-8504.1972.tb00192.x.

Urquhart, A., & Weir, C. (1998). *Reading in second language: Process, product and practice*. New York: Longman.

Vacca, R. T., & Vacca. (2008). *Content area reading: Literacy and learning across the curriculum* (9th ed.). Boston: Person education, Inc.

Weir, C., Hughes, A., & Porter, D. (1990). Reading skills: Hierarchies, implicational relationships and identifiability. *Reading in a Foreign Language, 7*(1), 505.

Weir, C., & Porter, D. (1994). The Multi-divisible or unitary nature of reading: The language tester between Scylla and Charybdis. *Reading in a Foreign language, 10*(2), 1-19.

Weir, C., Huizhong, Y., & Yan, J. (2002). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge: Cambridge University Press.

Wright, B. D., Linacre, J. M., Gustafson, J., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch measurement transactions, 8*(3), 370.