APPLICATION OF RASCH MODEL IN MEASURING THE QUALITY OF HEALTH AND WELLNESS FINAL EXAM QUESTIONS

Norlia Goolamally
Open University Malaysia
nolee@oum.edu.my

Suhaila Abdul Hamid
Open University Malaysia
suhaila_abdulhamid@oum.edu.my

Ahmadzulkarnain Ramli
Open University Malaysia
zulkarnain@oum.edu.my

Raijah A Rahim
Open University Malaysia
raijah@oum.edu.my

ABSTRACT

The assessment of the Health and Wellness module includes two components, 60% for the continuous examination and 40% for the final examination paper. In this paper we discussed the performance of students in the final examination paper which comprises 40 MCQs (MCQ). The 40 MCQ were constructed based on the Table of Test Specification developed by the subject matter experts at the faculty. The difficulty level of the questions was based on the 5:3:2 proportions. This ratio produces 20 MCQ at lower difficulty level, 12 MCQ at moderately difficult and 8 MCQ at higher difficulty. Results of 764 students were analysed based on their ability and item difficulty of the 40 MCQ using RaschWinsteps 4.01 software. Person Item Map Distribution (PIDM) showed that the item measure is between -3.62 to +4.97 logits and the person measure is between -1.28 to +4.09 logits. This finding indicated that while person ability exceeded item difficulty at the lower logit continuum of the Rasch model, at the upper logit continuum, some items are at higher difficulty level compared to students' ability. Item difficulty and separation indices statistics for both person and item are also of concern in this study. To ensure and enhance reliability and validity of the questions, it is highly recommended that the subject matter experts review the questions before applying it again in another exam or depositing it in the item bank system.

Keywords: *Assessment, Rasch Measurement Model, Item Bank, Item Difficulty, Student Ability*

INTRODUCTION

Test is an important tool in teaching and learning to determine the level of achievement and mastery of knowledge by the students. In addition, it can also be used to improve the quality of teaching and learning. Hence, student's assessment process should be taken seriously to achieve the programme learning outcomes and programme educational objectives.

MPU3313 Health & Wellness, a 3 credit-hour course, is one of the MPU courses offered by Open University Malaysia. This module consists of eight topics and the objectives of this module are to introduce the learners to the basic concepts of health and wellness, characteristics of a healthy lifestyle and personal responsibility to maintain health and wellness. This module helps learners to plan the strategies for emotional wellness, achieve spiritual support as well as recognise the importance of intellectual dimension in health and wellness. The course is delivered through a blended mode of teaching and learning as follows: (1) face to face; (2) online learning; and (3) self-managed learning.

For the final examination grades, MCQs are used to assess the achievement of the learning outcomes. A test using MCQs is an efficient and effective way to assess a wide range of knowledge, skills, abilities and attitudes (Haladyna et. al, 2002). The most important part of preparing multiple choice tests is to construct good questions. The real test of test designers is that the test provides a channel to justify students have achieved the required learning outcome. Test developers are not not only required to master the content but also to have a good understanding of the objectives of the assessment, besides having good skills in writing the items (Anna Siri Michela Freddan, 2011).

For this study, Rasch analysis was used to review how well the MCQs were constructed for the Health and Wellness module using Rasch Model. Rasch analysis is a statistical technique used in education to measure abstract constructs (A. Obermeier, 2009). The purpose of this study is to enhance students' success with suitable sets of questions. An accurate method of analysis of the questions needs to be properly performed in order to construct questions suitable for students' level of thinking.

LITERATURE REVIEW

A common feature of all universities is to produce students who are successful and are able to practice what they have learned to be of assistance to their communities and country. Student achievement is a major concern for all educators, students, universities and other relevant parties. Students' understanding of a subject is the most important component of learning. Educators should provide a level of assessment to commensurate student's cognitive level.

A good examination is one method to measure students' performance. Typically, the MCQs consist of incomplete questions or statements, referred to as a stem, and a set of two or more options comprising the plausible answers to the question. The student's task is to choose an option that gives the best answer to the question posed. The best answer is referred to as the main choice and the rest are called distracters. Obviously only one option is correct. Developing MCQs is not an easy task. It becomes very complex to form distracters that fit their purpose. In fact, the appropriate quality of MCQ is based on the availability of distracters. A good distracter should be able to discriminate between the informed and the uninformed student. (Kennedy Q., et. al, 2017). Therefore, item analysis will be used to assess the quality of these MCQs.

However, in order to accomplish this task of producing a good quality test items substantial amount of time, effort, and energy is required. In addition to the capability in enhancing subject matter knowledge, teachers need to spend relentless hours for tests planning including the process of assembling items, writing, and determining the difficulty level of each item (Richichi, 1996; Ahmad Zamri, 2010). This perhaps serve as one the possible reasons to explain why most test items developed by teacher are considered fail to discriminate between high and low ability; and items are not function according to their intended used (Richichi, 1996). In addressing this pitfall, literature suggests that much of the burden of test construction can be reduced when a large collection of good quality of test items is available to either teachers or test developers (Millman & Arter, 1984).

A large collection of test items along with their measurement characteristics is termed as item bank or interchangeably recognized as item pools. Item bank is a notion that is widely defined, from a loose and unrestricted definition such as 'any collection of test questions' (Millman & Arter, 1984) to 'a

composition of coordinated questions that develop, define, and quantify a common theme and thus provide an operational definition of a variable' (Wright, 1984). On the other hand, in a more detail manner, Choppin (1981), Rudner (1998) refer item bank as a large collection of good test items for which their quality is analyzed and known, and which are systematically stored in a computer so that they are accessible to students and teachers for measuring their achievement or ability.

Intriguing, Rudner (1998) highlights the application of item bank, namely, item banking provides substantial savings of time and energy in developing test due to its convenience in depositing and storing items.

There are other two main reasons to explain the sufficient application of item bank.   First, items in an item bank can be edited, withdrawn, and deposited (Rudner, 1998). Secondly, its capability to develop several test with the characteristics of flexibility, security, and consistency. The reasons mentioned convince item banking to be a useful tool for educational system to monitor educational achievement (MacCann & Stanley, 2009).

Considering what have been mentioned above, a well-developed item bank according to Educational Testing Service (1992) is only indicative through good quality of test items as listed below:

(i)   Items must be fair, valid and reliable in order to create fair, valid and reliable tests;

(ii)   A test is only as good as each item on it;

(iii)   If items don't really measure the standard, the test results will not be useful;

(iv)   Instructional decisions should be made on the basis of valid assessments;

(v)   Teachers need good data to make better use of their limited instructional time; and

(vi)   Older and repurposed items often do not address today's instructional standards.

As such, an item bank involves a thorough process of items filtering and calibrating before item storing (Wright & Bell, 1984) in developing good quality of items. Statistically, calibrating items means items require to be standardized for the purpose of precision. In other words, item bank should be developed and validated with accurate assessment in item level in order to produce a set of good quality items. Remarkably, a robust calibrated item bank that utilises Rasch Model provides numerous advantages to test developers such as flexibility, consistency, economy and security (Umar, 1999).

Advancement in using sophisticated computer software further enhances the possibilities of development of a calibrated item bank. Computer program such as WinSteps for test development and FastTest for item banking facilitates the development item bank where evaluation of each item and the formation of each test can be made with ease.

A variety area of studies have successfully developed and validated their test using Rasch Model. For example, Kazeem (1988) uses Rasch Model to construct achievement test while Abdel Fateh El Korashy (1995) utilizes Rasch Model to select items for mental test. Meanwhile, Barman and Boone (1997), Ludlow (2001) and Salzberger (2002) use Rasch Model to validate education tests while Heinemann et al. (1997) validate the measurement of stroke scale in medical research. The other significant empirical studies includes Baylor, Yorkston, Miller, and Amtamann (2009), Forkman, Boecker, Wirtz, Eberle, Westhofen, Schauerte, Mischke, and Norra (2009), Gothwal, Wright, Lamourex, and Pesudov (2009), Heinz, Smith, and Kassel, and (2009), as well as Muis, Winne, and Edwards (2009).

A well-constructed item bank enables teachers to design the best possible test for every purpose. Teachers can tailor each test to their immediate educational objectives and consider who is to be measured without losing contact with the common core of bank items. This is because it is not necessary for every student to take the same test in order to be able to compare results. Students can take the selections of bank items most appropriate to their levels of development. The number of items, their

level and range of difficulty, and their type and content can be determined for each student individually, without losing the comparability provided by standardized tests. Comparability is maintained because any test formed from bank items, on which a student manifests a valid pattern of performance, is automatically equated, through the calibration of its items onto the bank, to every other test that has been or might be so formed.

This study is an important step in creating OUM's item bank. However, additional analysis need to be carried out with other test booklets and test linking utilising common items in order to fit this purpose. Overall, it is important to realize that a well-planned and well-documented item bank is necessary for ensuring that the tests are fair, appropriate, reliable and valid. Rasch measurement model may contribute to this process.

Rasch Measurement Model

Rasch analysis is based on a stochastic or probabilistic model where Rasch measurement takes into account two parameters-tests namely the item difficulty and the person ability. These parameters are assumed interdependent. However, separation between the two parameters is also assumed. For example, the items (questions) within a test are hierarchically ordered in terms of their difficulty and concurrently, persons are hierarchically ordered in terms of their ability. The separation is achieved by using a probabilistic approach in which a person's raw score on a test is converted into a success-to-failure ratio and then into the logarithmic odds that the person will correctly answer the items. This is represented in a logit scale. When this is estimated for all persons, the logits can be plotted on one scale.

The items within the test can be treated in a similar manner by examining the proportion of items answered incorrectly and converting this ratio into the logarithmic odds of the item being incorrectly answered. A person's logit score can then be used as an estimate of that person's ability and the item logit score can then be used as an estimate of that item difficulty

$$P_{ni}(x_{ni} = 1/B_n, D_i) = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

ensuring the standards and objectives evaluation of student performance, item analysis is a process involved in assessing the quality and quantity of both items and tests as a whole through answers given by students (Anna Siri & Michela Freddano, 2011). The main purpose of item analysis is to improve the quality of the test by revising or eliminating ineffective items. It can provide diagnostic information about what students have learned and what they have not learned.

There are many different procedures for determining item analysis. The procedures used to assess the effectiveness of an item depend on the extent to which the researcher's priorities and for testing purposes. (K. Chellamani, 2013). In conclusion, item analysis is very meaningful when it comes to improving the teaching skills in the construction of tests items so that the questions can really measure the extent of the students' achievement in a subject.

Assessment is an important aspect of the teaching and learning process that aims to collect, interpret and analyze student performance. Quality of learning is determined by the quality of the assessment. Hence, using appropriate and quality assessment strategies enable students to be engaged in their own learning in a challenging but enabling environment. (Babar Khan, 2012)

Multiple-choice testing is easy to score and reduce the reliance on skills of writing and self-expression and become a great help for those students who have language problem. In addition, student may spend less time studying for the test (Kulhavey, Dyer, & Silver, 1975), and they take notes on different material compared to students expecting an essay exam (Rickards & Friedman, 1978) as cited in Henry L. Roediger III (2005).

In preparing the single best option type of multiple-choice questions (MCQs), it is recommended to write the options first. The possible options generated must be homogeneous based on the selected topic and context. The options should be readily understood and as short as possible. A good distracter should be inferior to the correct answer but should also be plausible to a non-competent candidate. Beside that a question stem is to be written with lead-in statement based on the selected correct option. The well-constructed MCQs should test the application of the context rather than just the recall of information so that it will enhance the problem solving ability. Quality MCQs must be free from any tricks and clues. (Mohammed O. Al-Rukban, 2006)

There are two major types of technical flaws that are commonly introduced by question writers namely; (1) flaws related to irrelevant difficulty - avoid things that may cause examinees to select an incorrect response; and (2) flaws related to examinees' test wishes - avoid responses that may help examinees select the correct answer.

The list of questions that can be used to guide the test developers in constructing good MCQs (Alfred Tenore, 2015) are shown in Table 1.

Table 1: Guide for MCQ Test Developers

Item	Stem	Lead-in	Option
• Has a single-best answer format been used? • Does it test application of knowledge rather than recall of isolated facts? • Does it satisfy the "cover test" rule? • Is it appropriately "balanced" (most of reading in stem & relatively short options)?	• Is the stem structured as a vignette, describing a specific situation?	• Is it phrased positively rather than negatively? • Does the phrasing avoid making the vignette irrelevant? • Is it focused so that it poses a clear question to be solved? • Is it structured as a complete sentence ending with a question mark?	• Are they homogeneous in content and phrasing? • Are they similar in length and parallel in structure? • Does each follow the lead-in both grammatically and logically? • Can they be rank-ordered on a single dimension (from most to least correct)? • Does the correct answer avoid repeating words used in the stem ("clang" clue)? • Are distracters phrased to avoid repetition that clues correct answer (convergence)? • Has the option-set been constructed to avoid "none of the above" "all of the above"? • Does each avoid the use of absolute terms (e.g. "always", "never")?

METHODOLOGY

This course is offered to 764 students in the January semester 2018. At the end of the semester, after twelve weeks of studying, they sat for the final examination in Health and Wellness. The study was conducted using the quantitative research approach and the following are the steps involved in the development of the final test items.

Step 1 – Development of the Test Specification Table
 The Test Specification Table was the first thing to be done before developing the items. This is a blue print to ensure that the items are equally distributed based on the difficulty level and student learning time. The purpose of this step is to state the content by topics and the number of student learning time (SLT) for each topic. At this stage the learning outcome is aligned to the content as shown in Table 2.

Step 2 – Instrument Development and Validation
 The instrument is in the form of a test consisting of 40 multiple choice items with four options A, B, C, and D. The items in the test cover the eight topics of the Health and Wellness module. The test items were self developed by the internal academicians.

Step 3 – Item Development
 The test developer is a lecturer at the university and has teaching experience for the Health and Wellness course. The items were developed using both English and Malay languages using backward translation from English language to Malay Language. The test items were developed based on the level of difficulty with the following criterion: Low, Intermediate, and High. Criterion Low includes Knowledge and Comprehension; Intermediate encompasses Application and Analysis and High covers Synthesis and Evaluation.

 Item content validation was conducted to ensure the items measure a specific learning outcome (Anastasi 1988). The validation and moderation was done at the faculty by another lecturer who is also the content expert of this course.

 The final test paper was distributed during the examination day to all students sitting for this course. Students are allowed 11/2 hours to complete the 40 MCQs.

Step 4 – Test and Item Calibration Design
 This study employs the Rasch Model software namely WINSTEPS 4.01 to model both students' ability and item difficulty.

The Rasch Model provides two infit statistics: infit and outfit Mean Square Statistics (MNSQ). The infit MNSQ is sensitive to unexpected responses to items near the person ability level and the outfit MNSQ is outlier sensitive. Mean square fit statistics are defined such that the model-specified uniform value of randomness is 1.0 (Wright & Stone, 1979). Person fit indicates the extent to which the person's performance is consistent with the way the items are used by the other respondents. Item fit indicates the extent to which the use of a particular item is consistent with the way the sample respondents have responded to the other items.

Rasch analysis provides reliability indices for both item and examinee's measure. For this analysis, values between 0.70 and 1.30 logits are considered acceptable (Bond & Fox, 2001). High reliability for both indices are desirable since they indicate a good replication if comparable items/examinees are employed.

Table 2: Mapping of Items to Learning Outcomes Based on Learning Domains for Health and Wellness

	Topics / Learning Outcomes	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Total	%*
		50% (20 Qs)		30% (12 Qs)		20% (8 Qs)			
1	**Concept of Health & Wellness**								
1.1	Describe the characteristics of healthy lifestyle		1		1				
1.2	Identify personal responsibilities	1					1		
1.3	identify factors characteristics of health & wellness			1				5	12.5
2	**Dimensions of Health & Wellness**								
2.1	plan physical activities that will contribute to PW		1						
2.2	Select strategies for balanced emotional wellness	1		1					
2.3	Employ spiritual support to enhance wellness		1						
2.4	Evaluate the significant of intellectual dimension				1			5	12.5
3	**Dimensions of health & wellness II**								
3.1	discuss strategies to enhance social wellness						1		
3.2	Evaluate factors that contribute to occupational wellness		1		1				
3.3	Discuss strategies to promote environmental wellness	1		1					
3.4	Interpret financial wellness		1					6	15
4	**Understanding Nutrition and Diet**								
4.1	Identify the component of healthy diet	1							
4.2	distinguish the dietary challenges for specific		1			1			
4.3	Compare healthy diet with therapeutic diet			1					
4.4	Plan for healthy diet		1		1			6	15
5	**Maintaining weight**								
5.1	Explain the concept of weight management		1						
5.2	Identify the causes of obesity					1			
5.3	Differentiate the two types of eating disorders	1		1					
5.4	Plan weight management therapy		1					5	12.5
6	**Preventing Diseases**								
6.1	Discuss the importance of disease prevention		1						
6.2	Discuss the risk factors, signs and symptoms of CVD					1			
6.3	Elaborate the types, risk factors and S & S of cancer	1		1					
6.4	Discuss the causes and risk factors for diabetes mellitus		1				1	6	15
7	**Financial health & Wellness**								
7.1	Differentiate financial health and financial wellness			1					
7.2	Identify the challenges confronting emerging dult in FW	1							
7.3	Discuss the elements of financial wellness					1			
7.4	Construct a financial fitness plan		1					4	10
8	**Medical and Health Insurance**								
8.1	Describe the different types of life insurance		1						
8.2	Explain the fundamental of insurance				1				
8.3	Select the appropriate medical and health insurance to suit the needs of the insured					1		3	7.5
Sub-Total		7	13	7	5	5	3	40	
TOTAL			20		12		8		100

FINDINGS AND DISCUSSION

Rasch model with the application of WINSTEPS 4.0.1, was used to analyse the data and to test the psychometric properties of the test in terms of validity and reliability of the instrument. This software enables the item difficulties to be ordered on the same linear scale along with the student measures of ability (Rasch, 1980).

The Person-Item Distribution Map (PIDM) details out the exact position of each student in relation to the respective item. Rasch Model tabulates the items (MCQ) on the right side and the persons are plotted on the left side of the map on the same logit scale which gives a precise overview on the student's achievement of each MCQ. This will give a clearer view of students' ability towards item difficulty.

In PIDM, item means, $Mean_{item}$ functions as a threshold value and it is set to zero on the logit scale. The higher the location of item from the $Mean_{item}$ the more difficult the item compared to an item on a lower location.

Similarly, with the person distribution, excellent students will be located at the top of the map while the low ability students will be placed at the bottom of the map. Therefore, the level of the person's ability can be identified from PIDM by looking at the separation between the person and item on the map. The bigger the separation means that the person is more likely to achieve the item.

The respondents' reliability index of 0.98 is a good value (Pallant, 2001) for the expected consistency on the logit scale for the answers on different sets of items that measure the same construct (Wright and Masters, 1982). Linacre (2007) stated that the reliability of respondents of ≥ 0.8 and respondents' separation index of ≥ 2.0 as good indices. The statistics generated by Rasch analysis estimate the degree of items suitability that measures latent variables, assuring the item-fit of the instrument are within an acceptable range.

Reliability and Separation Index

Table 3 shows the item reliability and separation index generated by Rasch analysis. The statistics shown indicate how Rasch model conform the item separation index and person separation index as well as the item reliability and person reliability.

The value of the item separation refers to the number of strata of item difficulties obtained in the questionnaire. As shown in Table 3 the value indicates that the items develop are well spread and the items are on the logits scale with high reliability. The value of the separation index for all respondents are align with the recommendations by Linacre (2005) which states that the separation value index of > 2.0 is good.

Table 3: Summary of Infit and Outfit MNSQ for Items and Persons

```
------------------------------------------------------------------------
|          TOTAL                      MODEL       INFIT        OUTFIT   |
|          SCORE    COUNT   MEASURE   S.E.    MNSQ   ZSTD   MNSQ   ZSTD |
|----------------------------------------------------------------------|
| MEAN     26.9     40.0     1.11     .41    1.00     .0    .99    .1   |
| P.SD      3.6      .1       .62     .03     .18    1.0    .56    .8   |
| S.SD      3.6      .1       .62     .03     .18    1.0    .56    .8   |
| MAX.     38.0     40.0     4.09     .85    1.79    3.6   6.76   4.4   |
| MIN.     12.0     39.0    -1.28     .39     .59   -2.7    .16   -1.5  |
|----------------------------------------------------------------------|
| REAL RMSE    .43 TRUE SD   .45  SEPARATION  1.04  Person RELIABILITY  .52 |
| MODEL RMSE   .42 TRUE SD   .46  SEPARATION  1.11  Person RELIABILITY  .55 |
| S.E. OF Person MEAN = .02                                            |
------------------------------------------------------------------------

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .52  SEM = 2.49

     SUMMARY OF 40 MEASURED Item
------------------------------------------------------------------------
|          TOTAL                      MODEL       INFIT        OUTFIT   |
|          SCORE    COUNT   MEASURE   S.E.    MNSQ   ZSTD   MNSQ   ZSTD |
|----------------------------------------------------------------------|
| MEAN    513.6    763.8      .00     .12    1.00     .0    .99   -.1   |
| P.SD    189.5      .4      1.71     .06     .05    1.8    .10    1.9  |
| S.SD    191.9      .4      1.73     .06     .05    1.8    .10    1.9  |
| MAX.    756.0    764.0     4.97     .36    1.11    3.7   1.23    3.5  |
| MIN.     19.0    763.0    -3.62     .08     .89   -5.3    .78   -5.2  |
|----------------------------------------------------------------------|
| REAL RMSE   .13 TRUE SD  1.71  SEPARATION 12.83  Item  RELIABILITY  .99 |
| MODEL RMSE  .13 TRUE SD  1.71  SEPARATION 12.89  Item  RELIABILITY  .99 |
| S.E. OF Item MEAN = .27                                              |
------------------------------------------------------------------------
```

The value for student separation is 1.04 which is relatively poor. This small value indicates that there is not enough differentiation among students to separate them into distinct performance level or strata. The analysis identified two groups of students only; good and weak. This value can be increased by widening the students' ability. The person reliability value of 0.52 is considered to be poor. The item summary provides the value of Separation G = 12.83. This value indicates that the items were sufficiently well separated in terms of difficulty. However, although the value of item separation conforms to replicability of items to comparable sample of examinees, there are other concerns; which are the poor and weak values of person separation and person reliability.

Person Item Distribution Map

The PIDM shows that Q26 is the most difficult item for students to achieve while the easiest item is Q7. There is a small separation between Q7, Q27 and Q2. This shows that the level of difficulty of the items is fairly spread out as there are gaps as we go up the logit scale. The gaps should be closed by introducing more items at different levels of difficulty so that students' performance level can be divided equally. Figure 1 shows that the Person mean value, $Mean_{person}$ for this analysis is 1.11 logits which is higher than the threshold item value, $Mean_{item} = 0$.

A total of 456 students (59.7%) were found to be above the $Mean_{person}$ and the highest person managed to score 4.09 logit. The achievement of the students shows that only 59.7 % have managed to achieve the learning outcome of this module. In contrast, 308(40.3 %) were located below the $Mean_{person}$ value of 1.11 logits.
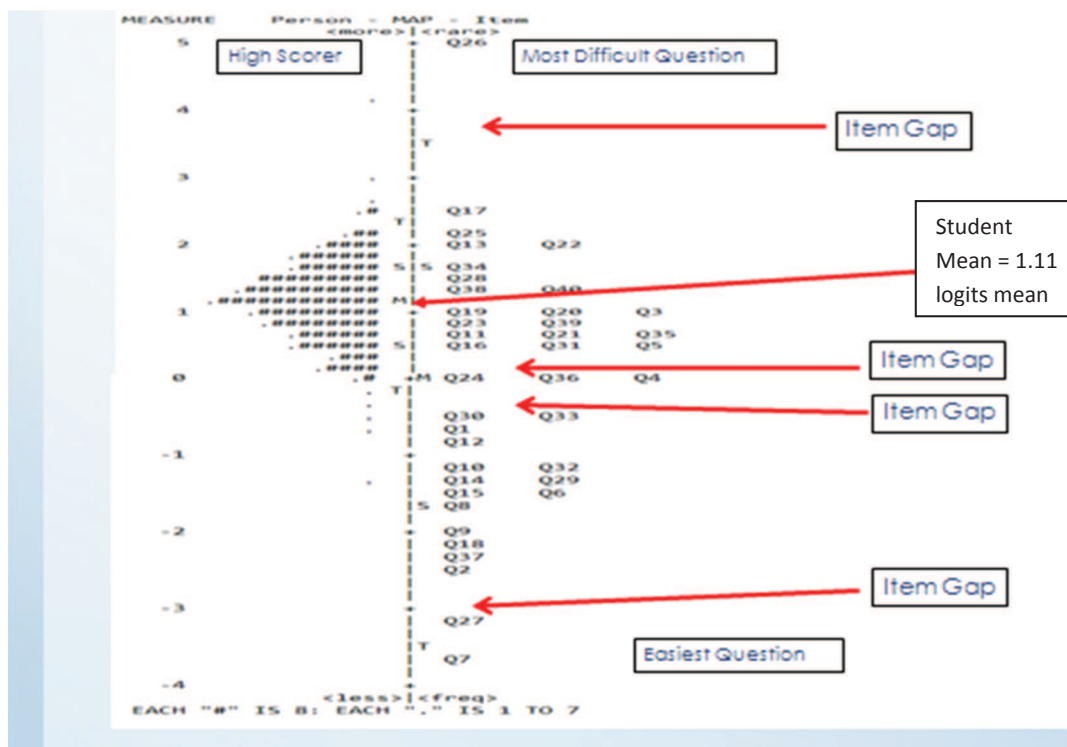

Figure 1: Person Item Distribution Map

For this exam paper, the item difficulties range from -3.62 logits to 4.97 logits. Students' mean value seems to be higher (M = 1.11 logits, SD = 0.41) when compared with item mean which is fix to 0.00 logits. The lowest score for person is -1.28 logits and it is higher than the lowest score for item which is -3.62 logits. There exists a separation gap of 2.34 logits. In this gap are eight items which are categorised as easy items and easily answered and passed by the lowest ability students.

The spread of the items can be calculated using the difference between Item Max and Item Min = 4.97-(-3.62) = 8.59 logit. Students will find more difficult to answer the question if the gap is wider. The largest gap is between Q26 and Q17. This means that item Q26 is a difficult question.

The distribution of persons based on their ability does not match with item difficulty. Items Q19, Q20 and Q3 are at the same difficulty level and measuring the same ability. Besides this example, Figure 1 shows other overlapping items which can be reduce d and improved to show a good spread of difficulty level and also person separation.

It is also observed that the ordering of item difficulty in the test paper is not following the sequencing of items. In best practices, the items are spread out based on their difficulty levels, from easy items at the beginning of the test paper to more difficult items toward the end of the paper. However, Figure 1 shows that items Q30, Q33, Q36, Q37, Q27 and Q29 which are supposed to be developed as difficult questions to test higher ability students are located below the $Mean_{item} = 0.0$. In contrast, items Q3, Q5 and Q4 which are supposed to test lower order ability is located above the $Mean_{item} = 0.0$.

Item Fit and Misfit

Statistical analyses for suitability of items were carried out to identify items that should be greater than 0.6 and less than 1.4 (Bond & Fox 2007). First, the fit statistic was performed on the outfit MNSQ then to the infit MNSQ statistics (Bond & Fox 2007).

Table 4 shows the items based on the infit MNSQ and outfit MNSQ statistics. The analysis shows the infit MNSQ and outfit MNSQ values for all items and respondents. The infit MNSQ and outfit MNSQ value of each item and the respondents should be in the range of 0.60 to 1.40 (Bond & Fox 2007) while according to Wright & Linance (1992), the total mean square infit and outfit mean square of each item and respondent must be located within 0.6 to 1.5.

The infit MNSQ values ranges from 0.89 to 1.11 and the outfit MNSQ values ranges from 0.78 to 1.23. These values are within the acceptable range of fit statistics from 0.70 to 1.30 (Bond & Fox, 2001). This indicate that all items in the exam paper are working together to define the student ability in the Health and Wellness final exam paper.

Table 4: Infit and Outfit Statistics of the 40 Health and Wellness MCQ

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
26	19	764	4.97	.23	.99	.0	.92	-.2	.14	.11	97.5	97.5	Q26
17	163	763	2.52	.09	1.11	2.1	1.23	3.3	.05	.25	77.9	78.9	Q17
25	201	764	2.22	.09	1.05	1.1	1.07	1.4	.18	.26	74.0	74.4	Q25
13	243	764	1.93	.08	1.00	-.1	1.01	.3	.27	.27	71.1	69.8	Q13
22	245	764	1.92	.08	1.02	.5	1.03	.7	.24	.27	68.2	69.6	Q22
34	298	763	1.59	.08	1.06	2.4	1.08	2.7	.18	.28	62.1	64.8	Q34
28	307	764	1.54	.08	1.03	1.1	1.03	1.0	.24	.28	62.8	64.3	Q28
38	332	764	1.39	.08	1.07	3.1	1.08	2.9	.17	.28	57.7	62.7	Q38
40	339	763	1.35	.08	1.05	2.4	1.06	2.2	.20	.28	57.9	62.4	Q40
3	389	764	1.07	.08	.99	-.7	.98	-.8	.30	.28	60.3	61.3	Q3
19	391	763	1.05	.08	.98	-1.1	.98	-.9	.31	.28	62.9	61.3	Q19
20	396	763	1.02	.08	.90	-5.3	.88	-5.2	.44	.28	68.5	61.4	Q20
39	428	764	.84	.08	1.08	3.7	1.09	3.5	.15	.28	56.5	62.3	Q39
23	443	764	.76	.08	1.04	1.8	1.04	1.3	.21	.27	59.2	62.9	Q23
35	452	764	.70	.08	.89	-4.9	.87	-4.5	.44	.27	70.5	63.5	Q35
21	461	764	.65	.08	1.07	2.9	1.08	2.5	.16	.27	59.9	64.0	Q21
11	471	764	.59	.08	.96	-1.7	.95	-1.7	.34	.27	66.8	64.7	Q11
16	476	764	.56	.08	.97	-1.0	.96	-1.2	.31	.27	64.8	65.1	Q16
31	490	764	.48	.08	.91	-3.3	.88	-3.6	.41	.27	69.2	66.2	Q31
5	496	763	.44	.08	.98	-.6	.97	-.8	.29	.26	68.2	66.8	Q5
24	551	764	.08	.08	1.00	.0	1.01	.2	.25	.25	73.2	72.5	Q24
4	553	764	.07	.08	.96	-.9	.93	-1.4	.31	.25	73.2	72.8	Q4
36	557	764	.04	.08	.94	-1.7	.87	-2.6	.36	.25	72.4	73.2	Q36
33	616	764	-.43	.09	.99	-.2	.96	-.5	.24	.22	80.8	80.7	Q33
30	619	764	-.45	.09	1.07	1.2	1.20	2.6	.08	.22	80.8	81.1	Q30
1	641	764	-.66	.10	.99	-.1	.97	-.4	.23	.21	84.0	83.9	Q1
12	651	764	-.77	.10	1.01	.2	1.03	.4	.17	.20	85.3	85.2	Q12
10	677	763	-1.09	.12	.99	-.1	.98	-.2	.20	.18	88.9	88.7	Q10
32	683	764	-1.16	.12	1.02	.2	1.08	.7	.13	.17	89.3	89.4	Q32
29	693	764	-1.31	.13	.97	-.2	.88	-.9	.22	.16	90.7	90.7	Q29
14	698	764	-1.40	.13	.96	-.4	.78	-1.7	.26	.16	91.4	91.4	Q14
6	702	764	-1.47	.13	1.00	.0	1.02	.2	.15	.16	91.9	91.9	Q6
15	707	764	-1.56	.14	.99	.0	.94	-.2	.17	.15	92.5	92.5	Q15
8	716	764	-1.75	.15	.97	-.2	.80	-1.3	.22	.14	93.7	93.7	Q8
9	727	764	-2.03	.17	.99	.0	.84	-.8	.17	.12	95.2	95.2	Q9
18	731	764	-2.15	.18	1.00	.0	1.03	.2	.12	.12	95.7	95.7	Q18
37	737	764	-2.37	.20	1.01	.1	.92	-.3	.10	.11	96.5	96.5	Q37
2	738	763	-2.44	.20	.99	.0	1.06	.3	.10	.10	96.7	96.7	Q2
27	751	764	-3.12	.28	1.00	.1	.89	-.3	.08	.08	98.3	98.3	Q27
7	756	764	-3.62	.36	1.01	.1	1.16	.5	.02	.06	99.0	99.0	Q7
MEAN	513.6	763.8	.00	.12	1.00	.0	.99	-.1			77.6	77.8	
P.SD	189.5	.4	1.71	.06	.05	1.8	.10	1.9			13.9	13.5	

Further analysis on the items validity can be carried out through the Point Measure Correlation (PMC) as shown in Table 4. In Rasch analysis, an item is considered to be misfit if all three controls (Point Measure Correlation, MNSQ and ZSTD) for the respective items are not in the range. The Point Measure Correlation denoted by x where $0.4 < x < 0.8$; the outfit mean square as y where $<0.5y<1.5$; and the outfit z-standard (ZSTD) as z where $-2< z<2$.

However, Table 4 shows that all items in this test fit the model. As an example, Item 17 has a Point Measure Correlation of 0.05, which falls outside the range of $0.4<x< 0.8$, so it is categorized as a suspected misfit item. When tested for the Outfit ZSTD = 3.3, it again falls out of the range and thus it becomes a highly suspected misfit item. However, for the last test, Outfit MNSQ = 1.23 falls within the range of $<0.5y<1.5$. Therefore, it is no longer categorized as a misfit item. This step has to be done for all items to determine if any of the questions should be labelled as a misfit item. The analysis shows that the items are considered to be in the acceptable range of the suggested fit indices.

However, to enhance the findings, Rasch measurement output allows the researcher to investigate the items based on the frequencies of responses of each of the distracters (Table 5).

The item, as the building block of a measure, must function in a way that contributes to the overall measure. When person reliability, item difficulty or person separation is not as expected as the findings in this research and not within the target range, a review on distracter performance has to be conducted (Livingston 2006). Complete item analysis is important during three stages, including item pre-testing, before scoring, and after scores have been reported (Livingstone 2006). This often provides a clue as to why an item may be too easy or too difficult. Similarly, when an item does not achieve the target level of discrimination, an examination of distracter functioning is important.

Table 5: Item Responses with Distracter Frequencies

Question	Data Code	Score Value	Data Count	%	Ability Mean		S.E. Mean	Infit MNSQ	Outfit MNSQ	PTMA Corr
					Mean	P. SD				
29	C	0	18	2	0.54	0.60	0.15	0.8	0.8	-0.14
	B	0	15	2	0.56	0.51	0.14	0.8	0.8	-0.13
	A	0	38	5	0.79	0.54	0.09	1.0	1.0	-0.12
	D	1	693	91	1.15	0.61	0.02	1.0	1.0	0.22
14	A	0	5	1	0.44	0.38	0.19	0.7	0.6	-0.09
	C	0	55	7	0.59	0.49	0.07	0.9	0.8	-0.23
	D	0	65	1	0.66	0.54	0.24	0.9	0.8	-0.07
	B	1	698	91	1.16	0.61	0.02	1.0	1.0	0.26
6	A	0	16	2	0.59	0.57	0.15	0.8	0.8	-0.12
	D	0	16	2	0.83	0.61	0.16	1.0	1.0	-0.07
	C	0	30	4	0.90	0.59	0.11	1.1	1.1	-0.07
	B	1	702	92	1.14	0.61	0.02	1.0	1.0	.015
15	C	0	2	0	0.55	0.08	0.08	0.9	0.7	-0.05
	B	0	14	2	0.64	0.71	0.20	0.9	0.9	-0.10
	D	0	41	5	0.79	0.51	0.08	1.0	1.0	-0.12
	A	1	707	93	1.14	0.61	0.02	1.0	1.0	0.17
8	C	0	15	2	0.50	0.33	0.09	0.8	0.7	-0.14
	B	0	25	3	0.60	0.61	0.12	0.8	0.8	-0.15
	D	0	8	1	0.71	0.49	0.18	1.0	0.9	0.22
	A	1	716	94	1.15	0.61	0.02	1.0	1.0	
9	A	0	11	1	0.44	0.26	0.08	0.8	0.6	-0.13
	C	0	23	3	0.70	0.51	0.11	1.0	0.9	-0.12
	B	0	3	0	0.79	0.47	0.33	1.1	1.0	-0.03
	C	1	727	95	1.14	0.62	0.02	1.0	1.0	0.17
18	A	0	30	4	0.70	0.63	0.12	0.9	1.0	-0.13
	D	0	3	0	1.32	0.29	0.20	1.9	1.5	0.02
	C	1	731	96	1.13	0.61	0.02	1.0	1.0	0.12
37	A	0	2	0	0.55	0.23	0.23	09	0.7	-0.05
	C	0	10	1	0.72	0.46	0.15	1.0	0.9	-0.07
	B	0	15	2	0.86	0.31	0.08	1.2	1.0	-0.06
	D	1	737	96	1.12	0.62	0.02	1.0	1.0	0.10
2	B	0	1	0	-0.14	0.90	0.22	0.5	0.3	-0.07
	C	0	13	2	0.77	0.77	0.20	0.9	1.1	-0.07
	A	0	11	01	0.84	0.62	0.02	1.1	1.1	-0.05
	D	1	738	97	1.12	0.61		1.0	1.0	0.10
	Missing*	*	1	0#	1.09	0.00				0.00
27	D	0	12	2	0.70	0.43	0.13	1.0	0.9	-0.08
	C	0	1	0	0.93	0.00		1.4	1.0	-0.01
	A	1	751	98	1.12	0.62	0.02	1.0	1.0	0.08
7	B	0	2	0	0.63	0.46	0.46	1.0	0.8	-0.04
	A	0	6	1	1.08	.043	0.19	1.5	1.3	0.00
	D	1	756	99	1.11	0.62	0.02	1.0	1.0	0.02

Table 6: Distracter Analysis for Item Q37

Item	Characteristics	Inference	Action
Q37	764 students answered 96% got D correct Options A, B and C are not attracting students to answer.	Distracters A, B and C not functioning. Students (low ability) marked the wrong answers.	• Q37 needs to be improved or removed. • Improved distracters to be plausible answer (to achieve level of discrimination)
	Location at -2.37 logit. This is below Meanitem = 0.0 This makes Q37 as an easy question	Based on question numbering, Q37 should be higher difficulty and testing higher ability.	

Similarly, the other items can be analysed in the same manner as shown in Table 6. In this way quality items will be deposited into the bank for future test application. This kind of analysis will increase test developers understanding in improving the function of the distracters. The role of distracters has become more salient as educator's demand instructionally relevant information. This means that we need information about what students know and can do as well as information about the misconceptions or errors in problem solving students continue to use.

CONCLUSION

This study has proved that students' performance on the Health and Wellness exam results can be measured using Rasch Measurement Model.

Furthermore, looking at the findings and discussions, we can also note that it was indicated that substantial number of questions from the forty MCQs requires significant improvements before they can be deposited into the Question Bank.

Before we proceed, let us take step back and ask ourselves, how do we qualify the questions before we store in the item bank? Can Rasch Measurement Model be useful as a tool for this? The answer is, YES.

From this analysis, there are several recommendations that can be proposed in order to increase the quality and reliability of the test items hence and later if implemented enables them (the questions) to be qualified and deposited into the item bank for later applications.

Item Gaps

Identifying, replacing and/or rephrasing the current items will enable the test developer to reduce the item gaps. Items should be constructed to test all levels of difficulty.

Order of Items Difficulty

Examining the responses of the distracters and summarizing the results of students in a table will assist test developers in rearranging the order of items to allow students to spend more time to answer tougher questions.

In the item bank system, the items have to be coded and tagged accordingly to reflect Student Learning Time (SLT), competency area and instructional objective as well the derived measures of item difficulty based on the Rasch analysis. Overlapping questions need to be revised and improved. The easiest items are placed at the bottom and gradually the most difficult items are on the top. Accordingly, the ordering of items should follow the same sequence from easy to difficult items.

Another strategy is reorientation in terms of item development will help. Asking item writers or editors to attend to the distracters requires explicit attention to the attractive aspect of the incorrect options. Asking ourselves: "Is this an effective attractor?" rather than "distracter" will improve our ability to be explicit about the intent of the option and characteristics of the students to whom it attracts. Improvements in item writing and item modifications that strive to make items and tests more widely accessible will contribute to the development of high-quality tests for wider audiences.

In conclusion, by using the Rasch measurement model, researchers have obtained high reliability value to the reliability test. This reliability test and respondents also indicates that the set of the questionnaire is valid and reliable to measure competitiveness. The findings have answered the possibilities designed to examine the suitability of the items in the competitiveness instrument. The item reliability is high and this means the item is stable.

REFERENCES

A. Obermeier Improving English test questions through Rasch Analysis, *Journal of Educational Research and Development* (2009) 107–113.

Abbdel Fateh EL Korashy (1995). Applying Rasch Model to the selection of items for mental ability test. *Educational and psychological measurement, 5*, 753–763.

Ahmad Zamri Kairani (2010). *Application of the bookmark method in setting performance standards in mathematics for form 2 students*. Unpublished PhD thesis. International Islamic University Malaysia.

Alfred Tenore, Danny G.P. Mathysen, Peter Mills, Mark Westwood, Jean-Baptiste Rouffet, Vassilios Papalois, Owen Sparrow, Zeev Goldik. (2015). *A guide to successfully writing MCQs: Executive summary*. European Council for Specialist Medical Assessments.

Anastasi, A. (1988). *Psychological testing*. 6th ed. New York: Macmillan.

Anna Siri & Michela Freddano. (2011). The use of item analysis for the improvement of objective examinations. *Procedia-Social and Behavioural Sciences, 29*, 188–197.

Baylor, C. R., Yorkston, K. M., Eadie, T. L., Miller, R., & Amtmann, D. (2008). The levels of speech usage: A self-report scale for describing how people use speech. *Journal of Medical Speech-Language Pathology, 16*(4), 191–198.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in human sciences*. New Jersey: Lawrence Album Associates.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental measurement in the human sciences*. ERL. Lawrence Erlbaum Inc. Publishers. London.

Boone, W. J. (1997). Science attitudes of selected middle school students in China: A preliminary investigation of similarities and differences as a function of gender. *School Science and Mathematics, 97*(2), 96–103.

C. Boopathiraj, & Dr. K. Chellamani. (2017). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research, 2* (2), 189–193.

Choppin, B. (1981). *Educational measurement and the item bank model*. In C. Lacey, & D. Lawton (Eds.), Issues in evaluation and accountability. London.

Forkmann, T., Boecker, M., Wqirtz, M., Eberle, N., Westhofen, M., Schauerte, P., Mischke, K., & Norra, C. (2009). Development and validation of the Rasch-based depression screening (DESC) using Rasch analysis and structural equation modeling, *Journal of Behaviour Therapy and Experimental Psychiatry, 40* (3), 468–478.

Gothwal, V. K., Wright, T. A., Lamoureux, E. L., & Pesudovs, K. (2009). Activities of daily vision scale? What do the subscale measures? *Investigating Ophthalmology & Visual Science, 51*(2), 694–700.

Heineman, A. W., Harvey, R. L., McGuire, J. R, Inberman, D., Lovell, L., Semik, & Roth, E. J. (1997). *Measurement properties of the NIH Stroke Scale during acute rehabilitation*. Retrieved Feb 18, 2007. http://stroke.ahajournals.org/cgi/content/28/6/1174

Heinz, A. J., Smith, E. V., & Kassel, J. D. (2009). Caffeine expectancy: Instrument development in the rasch measurement framework. *Psychology of Addictive Behaviors, 23*, 500–511.

Henry L. Roediger III, & Elizabeth J. Marsh. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology, 31*(5), 1155–1159.

Kazeem, A. M. (1988). *Application of Rasch model in the construction of achievement test in psychology*. Kuwait City, Kuwait. University Publication.

Kennedy Quaigrain, & Ato Kwamina Arhin. (2017). *Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation*. Educational Assessment & Evaluation. Retrieved at https://www.cogentoa.com/article/10.1080/2331186X.2017.1301013

Khan, B. (2012). Relationship between assessment and students' learning. *International Journal of Social Sciences and Education, 2*(1), 576–588.

Linacre, J. M. (2005). *Test validity, and Rasch measurement: Construct, content, etc. Rasch measurement transactions*. Retrieved on July, 2011 at http://www.rasch.org/rmt/rmt162fhtm

Livingston, S. A. (2006). *Item analysis*. In Downing, S. M., Haladyna, T. M. (Eds.), Handbook of test development (pp. 421–441). Mahwah, NJ: Lawrence Erlbaum.

Ludlow, L. H., & Mahalik, J. R. (2001). Congruence between a Theoretical Continuum of Masculinity and the Rasch Model: Examining the conformity to masculine norms inventory. *Journal of Applied Measurement, 2*(3), 205–226.

MacCann, R. G., & Stanley, G. (2009) Item banking with embedded standards. *Practical Assessment Research & Evaluation, 14*(17), available online at http://pareonline.net

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 1*(4), 315–330.

Ministry of Education Malaysia. (2004). *Integrated curriculum for Secondary School syllabus (Mathematics)*. Curriculum Development Center.

Mohammed O. Al-Rukban. (2006). Guidelines for the construction of MCQs tests. *Journal of Family Community Med., 13*(3): 125–133.

Muis, K. R., Winnie, P. H., & Edwards, O. V. (2009). Modern psychometrics for assessing achievement goal orientation: A Rasch analysis. *British Journal of Educational Psychology, 79*(3), 547–576.

Pallant, J. (2001). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows*. (Ver.10). Illinois: Allen & Unwin.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Richichi, R. (1996, October). *An item response theory analysis of multiple-choice items chosen at random from a Publisher's Test Bank*. Presentation at the 27th annual North Eastern Educational Research Association Conference, Ellenville.

Rudner, L. (1998a). *Item banking*. USA: Eric Clearinghouse on Assessment and Evaluation, Washington, DC.

Salzberger, T. (2002). *The illusion of measurement: Rasch vesus 2-PL*. Rasch Measurement Transactions, page 882.

Thomas M. Haladyna, Steven M. Downing, & Michael C. Rodriguez. (2002). A review of multiple choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Umar, J. (1999). *Item banking*. In G. N. Masters, & J. P. Keeves (Eds.), Advances in Measurement in Educational Research and Assessment. New York: Pergamon Press.

Wright B. & Linacre J. 1992. Combining and splitting categories. *Rasch Measurement transactions, 6*, 233–235.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 281–288.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, Why, How. *Journal of Educational Measurement, 21*, 331–345.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, ILL: MESA Press.