IS 3PL ITEM RESPONSE THEORY AN APPROPRIATE MODEL FOR DICHOTOMOUS ITEM ANALYSIS OF THE ANATOMY & PHYSIOLOGY FINAL EXAMINATION

[1]Hishamuddin Ahmad, [2]Siti Eshah Mokshein
[1,2]Faculty of Education and Human Development, Universiti Pendidikan Sultan Idris,
35900, Tanjong Malim, Perak, Malaysia

Abstract

Item response theory (IRT) offers some advantages over classical test theory and has been widely used to analyze dichotomous types of data in educational testing. This study aims to explore which is the most appropriate model to be used in the analysis of dichotomous items of the Anatomy and Physiology course. The study involved 971 nursing students studying in the Ministry of Health Malaysia training colleges. Exploratory factor analysis was performed on the data of the final examination paper containing 40 multiple-choice items. Results of the analysis showed that the unidimensionality and local independence assumptions were met. Data calibration was performed using an IRT-based software, *Xcalibre* based on the negative twice the log-likelihood statistic (-2LL). Results showed that the 3PL model is the most appropriate model for analyzing the data of the study. This study concludes that the 3PL model should be given a priority in analyzing the dichotomously scored items that involve guessing elements.

Keywords *Item response theory, classical test theory, dichotomous, Anatomy and Physiology, nursing program, Xcalibre, and 3PL model.*

Abstrak

Teori respons item (TRI) menawarkan beberapa kelebihan berbanding teori ujian klasik dan telah digunakan secara meluas untuk menganalisis data jenis dikotomus dalam ujian pendidikan. Kajian ini bertujuan untuk meneroka model mana yang paling sesuai untuk digunakan dalam analisis item-item dikotomus bagi kursus Anatomi dan Fisiologi. Kajian ini melibatkan 971 pelajar kejururawatan yang belajar di institusi latihan Kementerian Kesihatan Malaysia. Analisis faktor eksploratori telah dilakukan ke atas data kertas peperiksaan akhir yang mengandungi 40 item aneka pilihan. Keputusan analisis menunjukkan bahawa andaian unidimensionaliti dan kebebasan setempat telah dipenuhi. Penentukuran data dilakukan dengan menggunakan perisian berasaskan TRI, *Xcalibre* berdasarkan statistik *negative twice the log-likelihood* (*-2LL*). Hasil kajian menunjukkan bahawa model 3PL adalah model yang paling sesuai untuk menganalisis data kajian.

Kajian ini menyimpulkan bahawa model 3PL perlu diberi keutamaan dalam menganalisis item-item diskor secara dikotomus yang melibatkan unsur-unsur meneka.

Kata kunci *Teori respons item, teori ujian klasik, dikotomus, Anatomi dan Fisiologi, program kejururawatan, Xcalibre, dan model 3PL.*

INTRODUCTION

One of the major purposes of examinations is to obtain some information in order to evaluate students' performance. Thus, it is extremely important that the examinations contain sets of items which are valid, reliable and of high quality. Item analysis is normally carried out to assess the quality of items in a test. This is also being practiced in the training colleges of Ministry of Health Malaysia for the courses offered in their educational programs.

The information obtained from reports on student performance especially in Anatomy and Physiology (A&P) dichotomous tests, however, is usually limited in scope and depth. Not much information regarding description of student ability and test characteristics can be found. This is because traditionally, the proficiency of individual examinees is reported in terms of number-correct scores (number of items answered correctly), whereas the ability for a group of examinees is reported based on the mean and standard deviation of the number-correct scores. One problem with this approach is that students with similar number of correct answers will obtain similar scores, even though they may demonstrate different response patterns (ie, correct answers on different items). Thus, even though they obtain similar scores, they may not actually possess similar ability.

On the other side, reports related to the quality of test items are usually limited to indices of item difficulty (proportion of correct answers on the item) and item discrimination. A key problem with such indices, however, is that they depend on the group of examinees being tested (which may vary from one group to another) and therefore, do not adequately reflect the measurement quality of the items.

Item analysis provides a way of measuring the quality of items and seeing how appropriate they were for the examinees and how well they measured their ability. Continuous changes in educational outcome measures demand the use of newer and psychometrically sound instruments that produce valid scores. In the past six decades, problems that occur in the analysis of examinees' ability and test characteristics using traditional or classical test theory (CTT) have been successfully addressed in the framework of item response theory (IRT) (Dimitrov & Shelestak, 2003). IRT requires radical rethinking of measurement in a model that is not linear as compared to CTT which is a linear model. Moreover, the familiar concept of 'error' and 'reliability' do not appear in standard IRT models (de Ayala, 2009). IRT which is also known as modern mental test theory or latent trait theory is a body of theory that describes the application of mathematical models that expresses the relationship between an individual's response to an item and the underlying latent trait, also called construct or ability (Figure 1). Under IRT, the term 'ability' connotes a latent trait (performance, proficiency, etc.) that underlies the responses of examinees on the items of an instrument. This latent trait is expressed as beta (β) or theta (θ) according to the model used and is a continuous unidimensional construct that explains the covariance among item responses. Examinees at higher levels of ability have a higher probability of responding to an item correctly whereas examinees at lower levels of ability have a lower probability of responding to an item correctly. In the scope of this paper, the latent trait is referred as ability.

In IRT, the units of the ability scale known as logits typically range from -4 to 4. They represent the natural logarithm of the odds for success on the test items. For example, if a person succeeds on 80 percent and fails on 20 percent of the test items, the odds ratio for the success on the test is 4/1 = 4. Thus, the ability score of this person is the natural logarithm of 4 (or ln 4), which is 1.39 (Dimitrov & Shelestak, 2003).

Another advantage of IRT is its capability to specify reliability specific to each examinee. While reliability in CTT is summarized by a single index that is applied equally to all examinees regardless of ability level, IRT has the flexibility to estimate reliability uniquely for each examinee. This information can be very useful when different items are administered to different individuals or when building test forms with cut-scores or ability standards such that the forms can be built to maximize precision (minimize error) around those points on the scale. In IRT concept, the test information function indicates how well each ability level is being estimated by the test. IRT methodology forms the foundation of many psychometric applications such as for item development, item analysis and item banking (de Ayala, 2009). Despite its benefits, the appropriate application of IRT uses stronger (harder to meet) statistical assumptions than CTT, and typically requires larger sample sizes than those needed for CTT (Baker, 2001).

With an appropriate IRT model, the ability level of an examinee can be accurately estimated with any set of items that measure ability. This is because for a given ability, estimates of item characteristics hold true regardless of the group being tested. It means, a group of respondents of low ability will produce the same ICCs as a group of high ability (Baker, 2001). Conversely, the characteristics of test items (e.g., difficulty, discrimination and guessing) are accurately evaluated with any sample of examinees. The item parameters are not dependent upon the ability level of the examinees responding to the item (Baker, 2001). Therefore, educators can benefit a great deal from IRT analysis of student performance or test development (Dimitrov & Shelestak, 2003).

In educational measurement model, dichotomous IRT-based models that had been applied in the item analysis are the Rasch model and the logistic model. The idea of the development of IRT from CTT has started in 1950s by Frederic Lord and others when they realized that a completely new perspective on measurement was required (de Ayala, 2009). However, most of the literature in the 1950s through the 1970s used the term 'latent trait theory' as it reflected the use of an underlying hypothetical variable. Embretson and Reise (2000) stated that in the United States, the beginning of IRT often referred to a classic book authored by Lord and Novick in 1968 entitled 'Statistical Theories of Mental Test Scores'. The book contains four chapters related to IRT written by Allan Birnbaum made IRT visible. For a short period, Lord in 1977 referred to the field as 'item characteristic curve theory' (Baker & Kim, 2004). In 1980, Baker and Kim (2004) found that Lord's theory has gained acceptance as the label name of 'item response theory' reflects the basic concepts involved, and became one of the dominant topics of study among measurement specialists. Since then, IRT has subsequently been introduced and widely used in the fields of education and psychological testing (Hambleton et al., 1991).

IRT Dichotomous Models

For dichotomously scored items, IRT is based on a model that specifies the probability of correct response as a function of ability. The trace line produced by the model is referred to as an item characteristic curve (ICC). An example of an IRT model for a dichotomously scored item is shown in Figure 1.

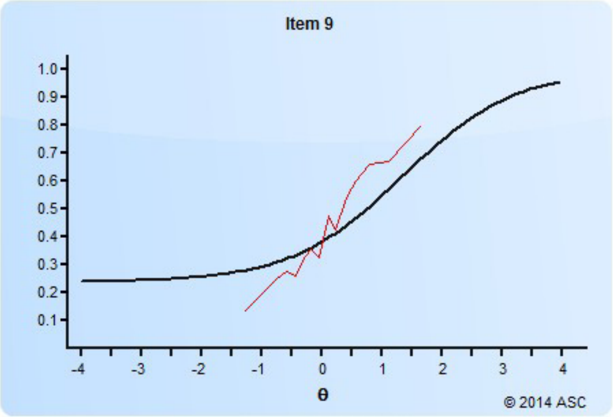

Figure 1 ICC for Item 9 from *Xcalibre* output

The most commonly encountered IRT models are for dichotomously scored items, in which item is scored ‘correct (1)’ or ‘incorrect (0)’. This would be the typical situation for multiple-choice items or multiple-choice questions (MCQ) that have a single correct answer and several incorrect options (distractors). There are several dichotomous IRT models which are based from Rasch model or logistic models. The Rasch model was introduced by a mathematician from Denmark named Georg Rasch (Azrilah, Mohd Saidfudin, & Azami, 2013). In 1952/1953, Rasch analyzed intelligence test data of Military Psychology Group an arrived at a model known as Rasch model for measuring examinee ability and item difficulty (Rao & Sinharay, 2007).

Other than the Rasch models, there are three most popular logistic models which vary with respect to the number of parameters they include to modify the shape of the ICC (Hambleton, Swaminathan, & Rogers, 1991). These three models are referred to as the one-parameter logistic model (1PL model), the two-parameter logistic model (2 PL model) and the three parameter logistic model (3PL model), so named because of the number of item parameter each model incorporates. As the number of parameters in the model increases (i.e., from 1 to 2 to 3), the model becomes more flexible, and thus can provide a more realistic reflection of how the expected response to each item is related to the underlying ability.

Both the Rasch and 1PL IRT model require that items have a constant value for discrimination, but allow the item to differ in their difficulty value. For the Rasch model, this constant is fixed to 1.0, whereas for the 1PL model, the constant discrimination value does not have to equal to 1.0 (de Ayala, 2009). In the 1PL models, however, items were assumed to have approximately the same discrimination parameter (CTB/McGraw-Hill, 2008; Meyer & Shi-Zhu, 2013). Mathematically, the Rasch and the 1PL model are equivalent. The 1PL model is the simplest item response theory model, and specifies an item's ICC using only one item parameter that reflects the difficulty of the item. For the Rasch model, the probability of a correct response is given by equation 1. The ability parameter is denoted by β and the difficulty parameter is denoted by *δ*. For 1PL model (equation 2), the ability parameter is denoted by θ and the difficulty parameter is denoted by the letter *b* (DeMars, 2010). As the value of *δ* or *b* increases, so too does the difficulty of the item. Notation *e* is a transcendental number whose value is 2.718 (Hambleton et al., 1991).

$$\mathrm{P}(\beta) = \frac{e^{(\beta-\delta_i)}}{1+e^{(\beta-\delta_i)}} \tag{1}$$

$$\mathrm{P}(\theta) = \frac{e^{a(\theta-b_i)}}{1+e^{a(\theta-b_i)}} \tag{2}$$

While the 1PL model has the advantage of simplicity, it lacks the flexibility to allow different items to have ICC of different slope (or steepness). The 2PL model (equation 3) overcomes this limitation of the 1PL model by including a second functional parameter (*a* parameter) that controls the steepness of the ICC. As a_i increases, so too does the steepness of the ICC. Because the steepness of the ICC reflects how well is the item ability to differentiate or discriminate, between individuals having different values of θ, *a* is commonly referred to as the discrimination parameter (Baker & Kim, 2004). Higher levels of item discrimination reflect a higher degree of information that the item provides about the respondent's ability level. As a result, the value of a_i is an indicator of how much information the item provides about the respondent's ability (de Ayala, 2009). Because the 2PL model considers how much information is provided by each item (via a_i), in estimating ability using the 2PL model different items are assigned different weights according to the item's value of a_i; the higher the value of a_i, the more weight is assigned to the item in estimating ability.

$$\mathrm{P}(\theta) = \frac{e^{a_i\,(\theta-b_i)}}{1+e^{a_i\,(\theta-b_i)}} \tag{3}$$

Item difficulty (*b*) and discrimination (*a*) are the two components of item analysis involved in 2PL model and 3PL model which are helpful in establishing the reliability of test scores (de Ayala, 2009). On top of *a* and *b* parameter, the other dichotomous IRT model which known as the 3PL model, includes the parameter *c* to represent the possibility of guessing (DeMars, 2010). The value of c_i reflects the lowest possible value of the item's ICC as ability becomes very low (also known as the lower asymptote of the ICC). Thus, if $c_i = 0.4$, then the probability of correct response for an individual with a very low level of ability would be 0.4. Because the value of c_i reflects the result of guessing behaviour, it is often referred to as the guessing parameter (de Ayala, 2009). The 3PL model specifies the probability of correct response on the *i*th item using equation 4.

$$\mathrm{P}(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1+e^{a_i(\theta - b_i)}} \tag{4}$$

For most MCQ tests, students with very low proficiencies have probabilities greater than zero of answering correctly even the most difficult items and usually through guessing. Guessing imparts a kind of unreliability to test score that is different from random measurement error, and this fact could result in statistical bias in analyses using number-correct scores (Chiu & Camilli, 2012). It means, any item which has a high possibility to be guessed by examinees have to be re-evaluated and further, should be considered to be reviewed or excluded. From a purely psychometric point of view, lower ability examinees who answer difficult items correctly are more likely (compared to higher ability examinees) to have guessed.

The Rasch model, 1PL and 2PL IRT assumes that no element of guessing are involved (c = 0) in examinees' responses. Since the events of guessing is actually a reality and the difficulty and discrimination estimates should be taken into account in providing information about the respondent's ability level, the 3PL model should be given priority in analyzing dichotomous item (CTB/McGraw-Hill, 2008). However, it is yet to be further examined whether or not the 3PL is the most appropriate model for the A&P data in this study. The use of an appropriate model for the data is crucial as it will affect the accuracy of the analysis and interpretations of the results.

OBJECTIVE

This concept paper seeks to explore whether or not the use of 3PL model of IRT is appropriate for analyzing the A&P final examination. It illustrates an IRT-based approach to perform analysis of dichotomous test items in the nursing program. However, while the typical IRT analysis is focused on the performance of individual items using a single model, this study deals with a selection of the best IRT model that fits the data.

METHODOLOGY

Sample

In early calibration of the IRT, Lord (1980) suggested a minimum sample size of 1000 to 2000 to be used for 3PL model. However, as time changes, and to address the problem of using a large sample size with IRT model, researchers have made further studies that showed recent computer software can provide acceptable calibration results by using smaller sample sizes. According to Guyer and Thompson (2011), *Xcalibre* is an available software that provides the most accurate and precise calibration, especially those involving the analysis of small sample size or a short test. In a study conducted by Guyer and Thompson, they found that *Xcalibre* is able to estimate the c parameter for dichotomous item well even with a sample size of 300. In addition, Guyer and Thompson also found that the accuracy of the parameter estimates increases with the increase in sample size for 3PL model, 2PL model, 1PL model and Rasch model.

The sample of this study consisted of 971 students enrolled in the diploma of nursing program at the training colleges of Ministry of Health, Malaysia. The A&P final examination was administered to students in the final examination of their first semester of January-June 2013 session.

Testing the Model Assumptions

In conventional statistics, ensuring a normal distribution assumption of data is one thing that often gives problems to researchers. Krylovas and Kosareva (2011) stated that the assumption of normality in real life is very seldom to be met. A more universal 'tools' (IRT) was created for the construction of the model even with a non-normal distribution data (variables / ability scores / parameters item). According to DeMars (2010), estimation procedure with IRT application does not require any assumption of a normal distribution of examinees' ability scores or normal distribution of item parameters. Since the assumption of a normal distribution of scores is not required in the application of IRT, the distribution of ability scores does not necessarily translate to a percentile score in the form of normality.

IRT models include a set of assumptions about the data to which the model is applied, namely unidimensionality and local independence (Embretson & Reise, 2000; Hambleton et al., 1991; He, 2010; Gao, 2011). The basic IRT models assume that the items analyzed measure a single dimension. Of course, defining single dimensions and writing items to capture them is challenging, and subfield experts sometimes disagree about the dimensions. The unidimensionality assumption which is common to the IRT models means that only one single ability is measured by the items that make up the test. This assumption is sometimes not met when cognitive, personality and test-taking factors might affect test performance. A few more factors that can affect the assumption are level of motivation, test anxiety, ability to work quickly and tendency to guess when in doubt about the answers. All these factors are said to contribute to random error. Thorpe and Favia (2012) found that, there is no one accepted method for determining unidimensionality. However, the unidimensionality of a scale can be evaluated by performing an item-level factor analysis, designed to evaluate the factor structure. Thus, dimensionality assumptions are sometimes tested with factor analysis prior to estimating IRT models (Jöreskog & Moustaki, 2001).

From the factor analysis, DeMars (2010) has suggested an eigenvalue analysis as a means to test the unidimensionality for dichotomous items. With this method, the eigenvalues are plotted in order, known as a scree plot. From the scree plot, a steep drop factors or 'elbow' followed by a sequence factors is to be inspected (Ruscio & Roche, 2012). If only there was a drop or dominant 'elbow' or bend in the scree slopes, then the assumption of unidimensionality is satisfied.

In this study, exploratory factor analysis using principal component analysis and extraction method with varimax rotation was performed on the data to assess the unidimensionality assumption. The data contain responses to 40 multiple choice items of A&P final examination from 971 samples. The scree plot was used as a guide in determining whether unidimensionality can be assumed (Figure 2). Results showed that unidimensionality has been concluded due to the presence of a dominant factor, (ie, a dominant 'elbow') as suggested by Ruscio and Roche (2012). In other way, we can see that the first eigenvalue was much greater than the others, suggesting that a unidimensional model is reasonable for this study data.

The second assumption of IRT models is that the items display local independence. This means that when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent. This is technically subsumed under the unidimensionality assumption and requires that, given their relationship to the underlying construct being measured is unidimensional, there is no additional systematic covariance among the items. In other words, local independence means that if the trait level is held constant, there should be no association among the item responses. Violation of this assumption may result in parameter estimates that are different from what they would be if the data were locally independent. The assumption of local independence can be violated for various reasons, including when items share similarities in their stems, wording, or reference passages (e.g., if respondents read several paragraphs and then respond to a series of questions about each, then the responses for each set of passages will be correlated (Wang & Wilson, 2005).

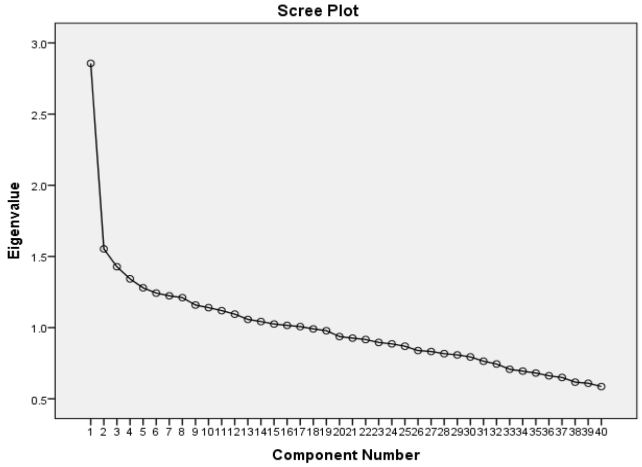

Figure 2 Scree plot for Eigenvalues

Unidimensionality is the most important assumption for all IRT models because when unidimensionality assumption is met, then the local independence is also obtained (Hambleton et al., 1991; Lord & Novick, 1968). Since the unidimensionality assumption of the latent trait measured in this study is considered reasonable, therefore the assumption of local independence is also met.

Research Instruments

The instrument used in this study was a set of multiple-choice items in the A&P final examinations. The paper consisted of 40 items designed to evaluate students' knowledge on nursing abilities as measured in the final examination of January-June 2013 session. All the 40 items are assembled according to a table of specification (Ministry of Health, 2003) and grouped into six subscales: Cells and Body Integration, Musculoskeletal System, Cardiovascular System, Respiratory System, Digestive System, and Intergumentary System (Ministry of Health, 2009). According to the Ministry of Health (Kementerian Kesihatan Malaysia, 2011), any MCQ examinations should contain items with different difficulty levels. The proportions should be approximately as the following: low level (20 to 40%), medium level (40 to 60%) and high level (20 to 40%). However, for the Diploma of Nursing program, the distribution of items according their difficulty levels in the A&P final examination was 30 percent of low level, 50 percent medium level, and 20 percent of high level (Ministry of Health, 2003).

Data Analysis Procedure

To select the most appropriate IRT model for the data, this study applied IRT-based software, *Xcalibre* to calibrate the data using the 1PL, 2PL and 3PL model. The data was not calibrated for the Rasch model using *Winstep* software because of the model's limitation to produce item discrimination parameter and guessing parameter. In order to choose the best model that fits well to the data, this study used a negative twice the log-likelihood statistic (-2LL or -2 times the log-likelihood) to evaluate the fit of models compared. According to Embretson and Reise (2000), the -2LL value of the data can be used to evaluate the fit of models compared. Comparing the values from different models can help decision which model represents a better fit.

Although there are many others IRT-based software packages that are free or commercially avaliable such as *Winsteps*, *Bigsteps*, *Noharm*, *Bilog-MG*, *Multilog*, and *Parscale* (de Ayala, 2009), *Xcalibre* software version 4.2 was used because of its ability to calibrate the data up to the 3PL model and its user-friendly features in graphical user interface (GUI) without Microsoft-Disk Operating System or MS-DOS command codes.

RESULTS

IRT Model Fit

In this study, -2LL statistic is used to test and compare the fit of the models, in which a smaller value indicates a better fit to the data (de Ayala, 2009; Embretson & Reise, 2000; Guyer & Thompson, 2013).

From the analysis shown in Table 1, this study found that -2LL value for the 3PL model was the smallest as compared to the 1PL and 2PL model. This indicates that the 3PL model provides a better fit to the data. In other words, the data in this study is more suitable to be analyzed using the 3PL model.

Table 1 -2LL results in comparing model-data fit

IRT models	1 PL	2 PL	3 PL
-2LL statistic values	42403	41951	41947

DISCUSSION

Generally, IRT is an appropriate model for item analysis as it offers several advantages over CTT. In the IRT models (Rasch, 1PL, 2PL and 3PL) discussed, the probability of a correct response is determined by the examinee's ability and item characteristics (difficulty, discrimination and guessing). The more elements that we can put into the model, the more information about the test and the ability of the test-takers can be described. Although the statistical assumptions and sample size of IRT are harder to be met compared to CTT, and typically requires larger sample sizes than those needed for CTT (Baker, 2001), the data and sample size of the examinees in this particular study sufficed the IRT assumptions.

This study also found that 3PL model is the most approriate model in calibrating the A&P final data. This finding was in accordance with the previous claim by CTB / McGraw-Hill (2008) that in a test that involved guessing, 3PL is a model that should be given priority in analyzing the test items. Moreover, the 3PL model is able to give more comprehensive information as compared to the other IRT models such as Rasch, 1PL, and 2PL.

CONCLUSION

This study provides a brief introduction to IRT and the related models. Some advantages of IRT over CTT are discussed. Basic concepts and selection the most appropriate model for the dichotomous data are described. The data examples in this paper illustrate only some of the functionality that IRT-based software, *Xcalibre* provides. Although, the Rasch, 1PL, 2PL and 3PL models are the most frequently used models in the applications, the 3PL model is usually the most appropriate model of choice in the calibration and analysis of dichotomous data involving guessing elements.

This paper only explains preliminary steps to be taken in the selection of the appropriate IRT model for item analysis. The actual item analysis will require more extensive analyses including measures of examinees' ability (theta), and item parameters (*a* parameter, *b* parameter an *c* parameter) and other input from contents.

REFERENCES

Azrilah, A., Mohd Saidfudin, M., & Azami, Z. (2013). *Asas model pengukuran Rasch: Pembentukan skala & struktur pengukuran.* Bangi: Universiti Kebangsaan Malaysia.

Baker, F. B. (2001). *The basic of item response theory* (2nd ed.). Wisconsin: ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker, Inc.

Chiu, T.-W., & Camilli, G. (2012). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*, 1-11.

CTB/McGraw-Hill. (2008). *Accuracy of the test scores: Why IRT models matter.* McGraw-Hill Companies Inc.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

DeMars, C. (2010). *Item response theory: Understanding statistic measurement.* New York: Oxford University Press, Inc.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Gao, S. (2011). *The exploration of the relationship between guessing and latent ability in IRT models.* Illinois: University at Carbondale.

Guyer, R., & Thompson, N. (2011). *Item response theory parameter recovery using Xcalibre™ 4.1.* Saint Paul, MN: Assessment Systems Corporation.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* USA: Sage Publications, Inc.

He, Q. (2010). *Estimating the reliability of composite scores.* Conventry, UK: The Office of Qualification and Examination Regulation.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 347-387.

Kementerian Kesihatan Malaysia. (2011). *Garis panduan penyediaan item anggota sains kesihatan bersekutu.* Putrajaya: Kementerian Kesihatan Malaysia.

Krylovas, A., & Kosareva, N. (2011). Item response theory applications for social phenomena modeling. *Societal Studies, 3*(1), 77–93.

Lord, F. M. (1980). *Application of item response theory to practical testing problem.* N.J., Erlbaum: Hillsdale.

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

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

Ministry of Health. (2003). *Diploma in nursing curriculum.* Kuala Lumpur: Training Division.

Ministry of Health. (2009). *Teacher's guide Year 1 Semester I: Diploma in nursing.* Kuala Lumpur: Ministry of Health.

Rao, C. R., & Sinharay, S. (2007). *Handbook of statistic 26.* Netherland: Elsevier.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*, 282-292.

Thorpe, G. L., & Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship, 20*, 1-33.

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random effects facet model. *Applied Psychological Measurement, 29*(4), 296-318.