ORIGINAL ARTICLE

Rasch Model Analysis of the Beck Depression Inventory-II among Malaysian School Students

Ahmad Zamri Khairani, Nor Shafrin Ahmad, Aziah Ismail, Rahimi Che Aman

School of Educational Studies, 11800 Universiti Sains Malaysia, Penang

ABSTRACT

Introduction: This study examines the psychometric characteristics of a translated version of the Beck Depression Inventory II (BDI – II) among Malaysian school students. **Methods:** The sample consisted of 257 boys and 302 girls. This study employed WINSTEPS 3.74 to provide statistics and other information from Rasch Model analysis, namely, the fit statistics, dimensionality analysis, rating scale analysis, reliability and separation indices, differential item functioning analysis, and distribution of items difficulty and students' ability. **Results:** Rating scale analysis showed that category 2 and category 3 of the ratings were not different. Meanwhile, Item 19 did not fit the model's expectations; and thus, it was omitted from further analyses. The scale demonstrated a high person reliability and a high person separation index. There were no items demonstrating gender DIF. The school students endorsed feeling guilty as the least severe symptom of depression, while committing suicide as the most serious symptom. **Conclusion:** In general, the BDI-II demonstrated acceptable properties in measuring depression symptoms among school students.

Keywords: Depression, Malaysia, psychometric, school students

Corresponding Author:
Ahmad Zamri Khairani, PhD
Email: ahmadzamri@usm.my
Tel: +604-6532965

INTRODUCTION

In Malaysia, findings from the National Health and Morbidity Survey (NHMS) revealed that 12 out of 100 students demonstrate mental health problem, with younger children showing a higher prevalence (1). Among the problems, depression is considered an important issue since it is seen as a global prevalent disorder (2). Signs of depression can be seen in cognitive, behavioral, and psychological aspects. From the cognitive aspect, signs of depression include feelings of hopelessness, negative view of oneself, and low self-esteem, while depressed mood, social withdrawal, and decline in personal appearance are some examples of signs related to behavior (3). Meanwhile, loss of appetite and difficulty in concentrating are signs of the psychological effects of depression (4). According to the Malaysian Psychiatric Association, depression may affect some 2.3 million Malaysians but it is remains undertreated (5).

In Malaysia, several instruments have been employed to measure depression such as the Beck Depression Inventory (BDI), the Depression, Anxiety and Stress Scale (DASS), Patient Health Questionnaire (PHQ), and Hospital Anxiety Depression Scale (HADS). The BDI remains as the most widely used instrument to measure depression (6). Nonetheless, this is not something unexpected since it is one of the most popular instruments worldwide (7). The BDI was first established by (8). The instrument underwent two major revisions in 1978, i.e. as the BDI-IA (9) and as the BDI-II in 1996 (10). The BDI-II identifies severity of 21 depression symptoms as follows: mood, pessimism, sense of failure, self-dissatisfaction, guilt, punishment, self-dislike, self-accusation, suicidal ideas, crying, irritability, social withdrawal, indecisiveness, body image change, work difficulty, insomnia, fatigability, loss of appetite, weight loss, somatic preoccupation, and loss of libido (10).

Assessment of depression using the BDI-II is abundance in Malaysia, with most studies focusing on samples with specific target group(s) for clinical interventions such as those suffering from urological problems (11), depressed patients (6) and postpartum women (12). The instrument has also been widely used in non-clinical settings such as among university students (13) and adolescents from single parents (14). Researchers in Malaysia also consider the quality of measurement from the BDI-II with a strong emphasis on the reliability and validity of the measurements. Psychometric evaluations of the BDI-II were mainly conducted using exploratory and confirmatory factor analysis, while evidence of concurrent, discriminant, specificity, and sensitivity are also presented (6).

Nevertheless, almost all the psychometric evaluations were conducted using the summative score procedure; i.e. the score for each items were summed to produce a

total score that was then compared to the standard cut-off scores to classify the respondents into the following categories: "normal", "sad", "mild", "moderate", "severe", and "extreme". However, this practice produces major flaws because it assumes that all items contribute equally to the measurement of depression symptoms even though it is highly likely that one symptom is more important than the other symptoms. In addition, as quoted by (15), the assumption of unidimensionality is routinely violated when measurements are conducted using ordinal data such as in surveys. The unidimensionality assumption requires that a test should measure only one unobserved construct at a time (16). According to (17), these flaws may lead to under- and over-estimated measurements of depression. Moreover, many studies have employed factor analysis techniques to determine the factor structure of the BDI-II. However, (18) cautioned for factor analysis is closely related to correlations between items for which skewed distribution scores might restrict the correlation range and produce inaccurate interpretation(s).

One possible source of the abovementioned problems is related to framework use in the measurement of depression. Literature shows that there are two distinct measurement frameworks as follows: The Classical Test Theory (CTT) and the Item Response Theory (IRT). CTT explains that the observed score (T) is a linear function of the true score (T) and the error score (E) specified by the formula X = T + E (19). However, researchers such as (20) have argued that the linear assumption is restrictive and not applicable when it comes to psychological constructs, because the assumption requires all items to have similar discriminations and locations on the ability scale. However, in procedures such as factor analysis, the items were assumed to have different discriminations. In contrast, the IRT employs a nonlinear relationship (called item characteristic curve, ICC) to explain the probability of observing a certain response with the latent trait. ICC provides flexible specifications on the relationship between the underlying trait and the items, independent from other test parameters such as response format, contexts, or theoretical assumptions about the response process (20). There are three types of IRT model available depending how many parameters are used to explain the relationship. In the one-parameter model, popularly known as the Rasch Model, the probability of observing a response is explained by the respondent's ability and the item's difficulty; whereas, in the two-parameter model, the relationship also includes item discrimination parameters. Meanwhile, the more complex three-parameter model, a pseudo-guessing parameter is added together with all other parameters mentioned earlier. While both two- and the three-parameter models are also widely used, the present study discusses measurement of depression within the framework of the Rasch Model.

The Rasch Model

Rasch model analysis helps to address the abovementioned limitations. The model has been widely used to the investigate psychometric characteristics of the instruments since the model provides users with a more comprehensive measurement framework compared to the summative score procedure. The Rasch Model is a family of modern test theory models that help to relate important parameters in the measurement of a construct, namely, the item difficulty and the respondent's ability parameters (19). In the Rasch Model, this relation is stated in the form of following equation (21):

$$P_{ni} = \frac{\exp(\beta_n = \delta_i)}{1+\exp(\beta_n - \delta_i)}$$

where,
P_{ni} = the probability of a respondent n correctly answered the item, i
β_n = ability of student n
δ_i = difficulty of item i

The process of estimating parameters is called calibration, and the score derived from the calibration is identified as a 'measure', and defined in logits unit. Measures from the Rasch calibration is essential for measurement because it possesses equal-interval properties of a person's ability and item difficulty. It should be noted that calibration does provides indications of differences between the two measures and evidence by how much they differed. With regards to the context of this study, a student with a higher ability measure demonstrates more positive responses (lower scores) towards depression symptoms compared to students with a lower ability measure. Similarly, items with higher difficulty measures are considered as more difficult to endorse (more high scores) than are items with lower difficulty measures. As such, the purpose of this study is to conduct the Rasch Model analysis of the BDI-II to provide more empirical information on the instrument using the Malaysian samples of high school students.

The Rasch Model is preferable when compared to other IRT models for several reasons. Firstly, since the model works with the fewest parameters, it is easier to work with. For example, the model can work with a smaller sample of respondents compared to the other models. Secondly, in the three-parameter model, all types of data are accepted since the model will adjust for any discrepancies in the data. However, the Rasch Model has rigid standards in controlling the types of data available. For example, erratic data that do not line up with the model's expectations will not be accepted for analysis. Guessing is also not accepted and is considered as reflecting the unreliability of the respondents, while discrimination variation is seen as a misleading item-bias interaction (22). As such, the quality of the data is already

assured by the model. Meanwhile, the 2-parameter model is often used in clinical health. The reason for this is that the model has a discrimination parameter, which indicates that each item has appropriate discriminatory power used to distinguish a patient's ability to perform an activity. However, for self-report instruments such as the BDI-II, each item must have the same discriminatory power so that the students' ability will not be influenced on the basis of how well they know the material being tested.

MATERIALS AND METHODS

Research Design

The present study adopted a cross-sectional study design. Data were collected in a single time period and the study involved translation and cultural adaptation of the original version, as well as provision of evidence of the psychometric properties of the BDI-II.

Sample of Study

Using the purposive sampling framework, a sample of 559 school students participated in this study. All samples were from five regular schools in the state of Penang using cluster sampling based on five districts as follows: North Seberang Perai, Central Seberang Perai, Southern Seberang Perai, Northeast Penang Island, and Southwest Penang Island. One school was selected from each district. The sample consisted of all form four students (average age = 16 years old). The use of sample with the same age is justified since age was also found to be important factor influencing measurement of depression (23). Table I presents the demographic characteristics of the sample according to their gender and ethnicity.

Table I: Sample of study

Demography	N	%
Gender		
Boys	257	45.89
Girls	302	54.11
Ethnicity		
Malay	436	78.00
Chinese	56	10.02
Indian	67	11.98

Instrument

The 21-item BDI-II was used in this study. Since the instrument was being tested among Malaysian high school students for the first time, a few modifications were made. Firstly, the instrument was translated into the Malay Language by a panel of experts consisting of a psychometric lecturer and a psychology lecturer using the back-to-back translation procedure. The panels translated the original version of the BDI-II, and then their translations were compared and revised to obtain the final translation draft. The draft was then examined by a language teacher with more than 10 years of experience to provide the final translated version. Secondly, Item 21 that relates to loss of libido such as "I have lost interest in sex completely" was excluded from the instrument based on the suggestions by school counsellors. This is because to teach school children about sex is considered a taboo (24). School teachers in Malaysia may delimitate the dissemination of information due to the traditional style of teacher–student relationship where the teachers are highly conservative and shy to even talk about sex and its related issues (24).

Procedure

The students were invited to fill in the paper-based questionnaire. Before that, they were explained as to the purpose of the study as well as how to respond to the questionnaire. We provided detail explanations on how to differentiate between the rating categories. We go through the first two items in the BDI-II with the students by reading aloud the items and explain what the items measured. We also asked the students questions such as what is the meaning of 'sad' (Item 1) and 'weak-willed' (Item 2) as well as giving examples to ensure that the students understood the meaning of each item. The students then proceeded with the rest of the items. The data were collected in their classroom during the teaching and learning process to ensure good responses. The students were able to complete the questionnaire within 15 minutes. The completed surveys were collected by the researcher and keyed in into an electronic database. Prior to filling in the questionnaire, the students were asked for their consent. Ethical clearance for this study was acquired from the university (USM/JEPem/17050263). Meanwhile, the researchers also obtained permissions from relevant stakeholders such as the Ministry of Education, the state education department, and principles from the respective schools. In addition, the researchers also followed the standard procedures throughout the study, especially in ensuring the confidentiality of the data.

Data Analysis

To assess the psychometric properties of the BDI-II, the Rasch model software WINSTEPS 3.74 (25) was used. The software employed the joint maximum likelihood algorithm to estimate item and person parameters from the measurement. In this algorithm, the available raw scores were sufficient statistics; therefore, there was no need for imputation or other treatment of the missing data. To achieve the objective measurement, the empirical data were subjected to several quality assurance analyses such as: (1) fit statistics, (2) dimensionality, (3) rating scale analysis, (4) reliability and separation, (5) differential item functioning (DIF), and (6) distribution of item's difficulty and student's ability. Table II presents the purpose and guiding criteria for each procedure.

RESULTS

Model-data Fit

The initial analysis showed that Item 19 (*0: I haven't*

Table II: Data analysis

Procedure	Purpose	Guidelines
Rating scale analysis	Whether the response categories function as intended	Minimum of 10 observations per category (26)
		The outfit mean-squares (MNSQ) value for each category is less than 2.0 (26)
		The values of average measures increased monotonically with increased category ratings (26)
		The values of threshold estimates increased with increased category ratings (26)
Model-data fit	To ensure the empirical data matches the model's specifications	The values of the infit and outfit MNSQ between 0.6 – 1.4 (21)
Unidimensionality	To examine whether the scale is measuring a unidimensional construct	The eigen value of the first contrast is less than 2.0 (27)
Reliability and separation indices	To examine consistency of the measurement	Reliability index of >.80 Item separation index of >2.0 (21)
DIF analysis	To investigate construct equivalence across groups	DIF contrast statistic < .5 logits (28).
Distribution of item difficulty and students ability	To provide evidence of matching between the samples and the items	Mean of item difficulty measure matches the mean of person ability (26)

lost much weight, if any, lately; 1: I have lost more 2.3 kg; 2: I have lost more than 4.5 kg; 3: I have lost more than 6.8 kg) demonstrated a misfit to the Rasch Model's expectations based on the high value of outfit MNSQ of 2.12. The result suggested that Item 19 was also considered problematic based on its low item loadings in the exploratory factor analysis procedure. Therefore, we proceeded with the analysis by removing Item 19 from the scale and then re-applying the Rasch analysis. Table 3 shows that all the remaining items demonstrated acceptable infit and outfit MNSQ values of between 0.6 to 1.4.

Dimensionality

The PCA of residuals showed that an eigenvalue of 1.7 in the first contrast was larger than the eigenvalues for the other contrasts of 1.4, 1.3, and 1.2, respectively. These values were less than 2.0, which gave an indication of no threat towards the secondary unintended dimension apart from depression in the measurement (27). Nevertheless, since the dimensionality assumption is an integral part in the Rasch Model analysis, we investigated this further because the data violated other indicators of dimensionality assumptions. For example, the raw variance explained by the measures is only 33.4%, which is less than the intended value of 40%. Also, the ratio between the raw variance explained by items and the unexplained variance in the 1st contrast was only 3.18, which is less than the cut-off value of 5. As such, we investigated the randomness of the standardized residuals. Figure 1 shows that there are eight items labelled A to H that were separated significantly from other items. Nevertheless, since all the items are measuring symptoms of depressions and no other unintended construct, we can conclude that the unidimensionality assumption was met.

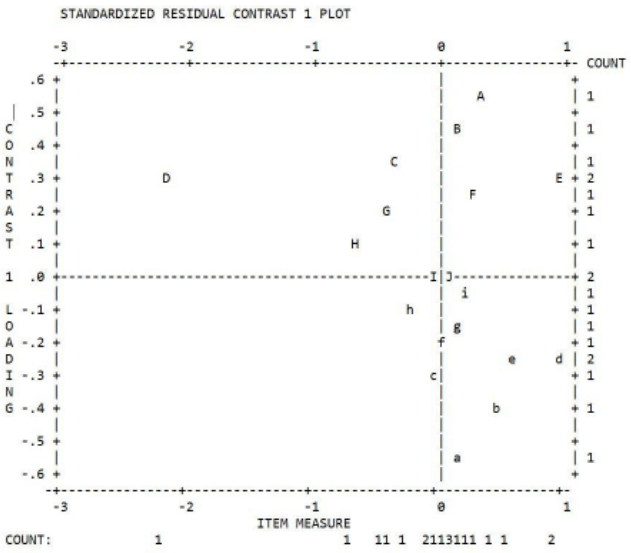

Figure 1: Residual contrast plot

Rating Scale Analysis

Initial results showed that the measurement of depression satisfied the first two criteria (the number of observations and the value of outfit MNSQ). However, for the third criterion, four items (Item 4, Item 9, Item 10, and Item 11) did not show monotonic advance of the average measures. For Item 4, the value of average measure for Category 3 was -.60, a decrease from -.35 in Category 2. Similar cases were also observed with regards to Item 9, Item 10, and Item 11, indicating that there were problems related to Category 2 and Category 3. As such, the rating scale needs to be revised, particularly by collapsing Category 2 and Category 3. Table 3 depicts that, after combining Category 2 and Category 3, the average measure for every item advanced monotonically with each category, as intended. It is noteworthy that satisfying the third criterion also resulted in the fulfilment of the fourth criterion. In general, there were three distinct peaks of the probability curve for each response category and all category frequencies displayed regular distributions. Nevertheless, the difference between category 0 and category 1 was only .56 logits whereas between category 1 and category 2 was only 1.12 logits, which were less than the intended value of 1.4 logits.
Insert Table III.

Reliability and Separation Indices

The reliability of item difficulty measures was high (.98), while the separation index was also high at 6.76. The high reliability might be due to the wide range of item difficulty in the scale. Both statistics are slightly higher compared to those measured before collapsing category 2 and category 3 (reliability = .97, separation = 5.87). Meanwhile, the students' depression reliability was .77, with a separation index of 1.81. Similar to the

Table III: Rating scale category statistics

Item	Category				Category		
	0	1	2	3	0	1	2+3
1	-1.75	-1.16	-.36	-.20	-1.42	-.77	.24
2	-1.68	-.93	-.39	-.03	-1.36	-.52	.19
3	-1.87	-.93	-.63	-.14	-1.55	-.51	-.14
4	-1.76	-.91	-.35	-.60*	-1.43	-.48	.03
5	-2.33	-1.51	-.70	-.41	-2.00	-1.17	-.15
6	-1.82	-1.21	-.74	-.35	-1.50	-.81	-.15
7	-1.86	-.95	-.20	-.01	-1.53	-.55	.44
8	-1.94	-1.14	-.79	-.47	-1.61	-.75	-.26
9	-1.62	-.71	.08	.10*	-1.28	-.26	.75
10	-1.73	-.75	-.23	-.63*	-1.37	-.28	-.18
11	-1.72	-.83	-.20	-.61*	-1.39	-.36	-.08
12	-1.59	-.92	-.62	-.58	-1.24	-.50	-.14
13	-1.66	-.81	-.54	.05	-1.33	-.38	.07
14	-1.62	-.87	-.52	-.29	-1.29	-.42	.09
15	-1.92	-1.11	-.55	-.02	-1.58	-.73	-.01
16	-1.58	-.83	-.43	-.39	-1.25	-.39	.11
17	-1.79	-1.10	-.46	-.28	-1.43	-.75	-.13
18	-1.56	-.70	-.16	-.12	-1.22	-.23	.48
19							
20	-1.63	-1.13	-.63	-.06	-1.29	-.74	.01

item statistics mentioned above the students' statistics are slightly higher when compared to before collapsing (reliability = .74, separation = 1.69). The statistics reported in this study were close to the intended values of .80 (reliability) and 2.0 (separation) respectively (21). This showed that there were enough items in the revised BDI-II to distinguish the students according to their severity of the depression symptoms.

DIF Analysis

DIF analysis was conducted to assess construct equivalence between male and female subgroups. The DIF contrast statistics (Table 4) show differences in the mean measures between the male and the female students. For example, since the item difficulty measure for Item 1 for the male students was estimated as .18 logits compared to -.19 logits for their female counterparts, the DIF contrast statistics was .18 – (-.19) = .37 logits. That is, the male students had more difficulty in agreeing to Item 1 compared to the female students. In contrast, the female students (measure = .32 logits) have more difficulty to agree with Item 2 compared the male students (measure = -.07 logits). As such, the DIF contrast statistics for Item 2 was calculated as -.07 – .32 = -.39 logits. In general, even though there are differences in item difficulty measures for the BDI-II (except for Item 8 and Item 9), these differences are not significant since all the values of the DIF Contrast are less than .5 logits.

Distribution of Items Difficulty and Students' Ability

Figure 2 shows the distribution of items difficulty and students' depression measures. The figure shows the BDI-II items were coded as D1 to D20, while the students

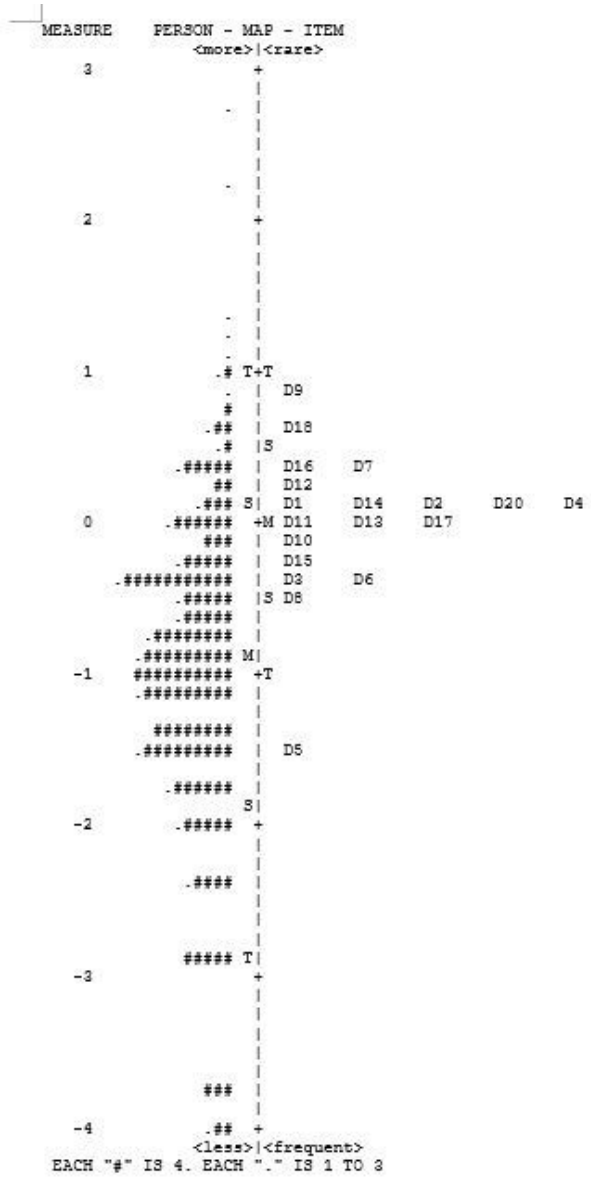

Figure 2: Distribution of item difficulty and student depression measures

were indicated by #, with each # representing 4 students and "." representing a proportion of 3 students. It showed that feeling guilty (D5: measure = -2.19 logits) was the easiest-to-agree item based on its lowest difficulty measure. That is, more students were answering category 0 (I don't feel particularly guilty) and category 1 (I feel guilty a good part of the time) compared to category 2 (I feel quite guilty most of the time) or category 3 (I feel guilty all of the time). In other words, feeling guilty was the least severe symptom of depression. They also endorsed that self-accusation (D8: measure = -.72 logits) was the second most severe depression syndrome, and this was followed by feeling of being punished (D6: measure = -.47 logits). In contrast, the students endorsed that committing suicide (D9: measure = .94 logits) was the most severe symptom of depression. Results also showed that loss of appetite (D18: measure = .91 logits) and insomnia (D16: measure = .58 logits) were among the most severe symptoms of depression.

In general, we can see that the mean score (indicated by the letter 'M") for the item measure (0.00 logits) was higher than the mean score of the students' measure (-.93 logits). This indicated that the students had difficulty to agree with the BDI-II items. As such, it can be inferred that the students had less severe depression symptoms. Another important observation was that there were several students placed at the upper end of the scale, indicating that they had higher depression symptoms compared to others. A further investigation revealed that 10 students had depression measures greater than the difficulty measure of Item 9. Nevertheless, we were not able to trace their identity because of the lack of information about the students. Also, there were a large number of students (N = 275 or 49.1%) with depression measures below difficulty measure of Item 8 that was targeted with only one item (Item 5). As such, the estimation of their depression measures might not be as accurate as that of other students who were targeted by many items at the upper parts of the scale.

DISCUSSION

The main finding from the present study was that the four-point rating category did not function well with the Malaysian sample of high school students. Instead, there is a need to combine category 2 with category 3 to achieve a better measurement of depression. For this purpose, we suggested dropping category 3 from the instrument. The reason behind this is that it is better not to include category 3 in the BDI-II because the sample of students did not suffer from severe depression. Nevertheless, the present finding does not fit well with the current views on depression among Malaysian school students. For example, newspapers and social media are quick to relate suicide incidents among students with depression. However, it should be noted that the literature has shown that apart from depression, there are various factors that influence suicide such as prior attempts, personality, family factors, specific life events-traits (such as relationship break-ups, the death of friends, and peer rejection), contagion-imitation, and availability of means (29). Hence, there is a need to see each and every incident more thoroughly. This includes specifying each incident in detail so that the cause(s) of the incident can be explained better. More importantly, stakeholders can learn from the incidents that happen so that similar events could be prevented.

The present study also does not agree well with other studies on depression that showed students experiencing higher prevalence of depression (30, 31). Nevertheless, it should be noted that these studies employed the method of calculating prevalence (i.e., the proportion of the population that shows syndrome of depression at a particular time). While the procedure is widely used in mental health studies, an important drawback of using prevalence is that the researchers need to ensure that the samples used are representative of the population so that the findings are generalizable. This is rather difficult to fulfil especially when there is a large number of school students (4 735 116) nationwide (32). As such, one might speculate that the sample of boarding school in the capital city of Kuala Lumpur (31), as well as in a district of Pasir Gudang, Johor (30), might not be representative. Therefore, the findings might not be generalized to the population of Malaysian school students.

The present study also reported that Item 19, which is related to weight symptom, does not fit the Rasch Model expectation based on the high value of the outfit MNSQ. That is, the item measured noise apart from weight syndrome. This noise caused the item difficulty estimated from low-ability persons to differ noticeably from the item difficulty estimated from high-ability persons (33). Nevertheless, (34) also explained that outfit problems are less of a threat to measurement and easier to manage by replacing responses of the person with missing values and reanalyzing changes to the item difficulty measures. In this study, we found that this practice causes significant changes in the item difficulty measures; and, thus, Item 19 was eliminated from further analysis. Note that deleting misfit items is considered a practical way to obtain better measurement in the Rasch framework. For example, in the study of (35), Item 19 and Item 21 were identified as misfits for the sample of 279 college students and were deleted from further analysis. However, we were not able to find other studies to draw an exhaustive conclusion to explain why this item shows misfit behavior towards the model's expectations. Nevertheless, we believe that this result can be traced back to the language issue. Even though the translation process has been conducted carefully, there is a probability of meaning loss when the item is translated to the Malay Language. The reason for this, as rightly observed by (36), item translation is a challenging task especially when it involves emotional and physical experiences. This is because the process does not only involve semantic, but also the cultural aspect of the translated language.

Apart from the two issues mentioned above, the present results support the adequacy of the BDI-II in measuring depression symptom among school students. For example, the present study agrees well with (35) to confirm unidimensionality property of the BDI-II. Also, the present study shows acceptable values of items' and students' reliability and separation indices that give indication of the consistency of the measurement of depression symptoms. The present finding is also significant in terms of interpretation of the BDI-II items among the boys and the girls. More specifically, boys and girls with the same depression intensity were found to have responded similarly to all of the items. As such, it is possible to interpret the results as intended. Studies such as (37) showed that several items may be understood differently by different genders. If such a

case, there is a need to study the items to identify the reason so that the interpretations is adequate for the sample of respondents.

This study also has important implications especially related to the utility of the BDI-II among students in Malaysia. According to the manual of BDI-II, screening for depression was conducted according to the following guidelines: scores 0–13 showed minimal depression, 14–19 (mild), 20–28 (moderate), and 29–63 (severe). However, this guide is considered inappropriate because as shown in this study, there is one item (Item 19) that is not suitable for use in the context of students in this country. Naturally, there is a need to set cut scores dedicated to BDI-II consisting of only 20 items so that at-risk students can be referred for treatment. However, this is not an easy task and requires in-depth study. In addition, new cut scores that are formed should also be tested for their suitability.

Perhaps it is too early to see whether the findings of this study can improve screening for depression. Nevertheless, it is certain that this study can facilitate school teachers and counsellors to identify early symptoms among students without clinical knowledge of depression. For example, teachers may flag students with easy-to-identify symptoms of depression such as social withdrawal, self-dislike, and self-dissatisfaction and investigate further. Early detection like this is known to be helpful for students to better deal with depression. The findings of the present study are bound by several limitations. Firstly, this study employed a relatively small sample of school students because of the difficulty in obtaining the permission from related stakeholders. Replication with a bigger and a more representative samples might be able to shed more light on this, particularly on the fit indices of Item 19 since more samples will provide more responses for a better estimation of its item difficulty. Also, a bigger sample size may be useful, particularly with regards to better DIF analysis (17), particularly since studies such as (38, 39) shows that gender can have impact on the response pattern of depressive symptoms, which is not observed in this study. Secondly, the present study focused heavily on the Rasch statistics to evaluate the BDI-II. As such, we exclude the information of the clinical parts of the instrument. Like any other screening test, one of the important purposes of the BDI-II was to classify the takers into normal, sad, mild, moderate, severe, and need serious treatment categories so that they can undergo appropriate treatment. As such, there is a need to set new cut scores for these classifications so that the students may benefit from the early detection of depression.

CONCLUSION

Assessment of psychometric properties of an instrument is important since the quality of the information gathered somewhat depends on the quality of the instrument used. In this article, we provide a brief explanation on the drawback of summated score strategy in measuring depression symptoms using the BDI-II. We also employed the Rasch model analysis to evaluate the instrument. The results showed that apart from the issues of the adequacy of the four-point rating scale and the misfit of Item 19, the BDI-II demonstrated acceptable properties in measuring depression symptoms. Also, the data explained that feeling guilty, self-accusation, and feeling of being punished were among the most severe symptoms of depression. As such, school counsellors may observe these symptoms among the students as an early measure to identify those with depression. This is important because even though depression is common, it is mostly unrecognized particularly due to the fact that most teachers have little knowledge in identifying symptoms of depression.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education under the Fundamental Research Grant Scheme [203.PGURU.6711548]. We also would like to thank Universiti Sains Malaysia for making this research possible.

REFERENCES

1. Institute for Public Health. National Health and Morbidity Survey 2015 (NHMS 2015). Vol. II: Non-Communicable Diseases, Risk Factors & Other Health Problems. Ministry of Health Malaysia. [Internet]. 2015 [cited 2019 June 1]. Available from http://www.moh.gov.my/moh/resources/nhmsreport2015vol2.pdf
2. Ferrari AJ, Somerville AJ, Baxter AJ, Norman R, Patten SB, Vos T, Whiteford HA. Global variation in the prevalence and incidence of major depressive disorder: a systematic review of the epidemiological literature. Psychological Medicine. 2013;43:471-81. doi: 10.1017/S0033291712001511
3. Huberty TJ. Best practices in school-based interventions for anxiety and depression: Best practices in school psychology V. In: Thomas A & Grimes J, editors. Bethesda: National Association of School Psychologists; 2008.
4. American Psychiatric Association. The diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Association Publishing. 2013. 1520 p.
5. Deva MP. Depressive illness–the need for a paradigm shift in its understanding and management. Medical Journal of Malaysia. 2006;61(1):4-6.
6. Mukhtar F, Oei TPS. A Review on the assessment and treatment for depression in Malaysia. Depression Research and Treatment. 2011;123642. doi:10.1155/2011/123642

7. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. 2006. 768 p.
8. Beck AT, Ward CH, Mendelson M, Mock JE. Erbaugh JK. An inventory for measuring depression. Archives of General Psychiatry. 1961;4:561-571.
9. Beck A, Rush AJ, Shaw BF, Emery G. Cognitive therapy of depression. 1979. 425 p.
10. Beck AT, Steer R A, Brown GK. (1996). BDI-II, Beck Depression Inventory Manual. 1996. 38 p.
11. Quek KF, Low WY, Razack AH, Loh CS. Beck Depression Inventory (BDI): A reliability and validity test in the Malaysian urological population. Medical Journal of Malaysia [Internet]. 2001 [cited 2019 May 19];56(3):285-292. Available from https://europepmc.org/abstract/med/11732072
12. Mahmud WMRW, Awang A, Herman I, Mohamed MN. Analysis of the psychometric properties of the Malay version of Beck Depression Inventory II (BDI-II) among postpartum women in Kedah, North West of Peninsular Malaysia. Malaysian Journal of Medical Sciences. 2004;11(2):19-25.
13. Zanil MF, Hashim NNWN, Azam H. Prediction of Beck Depression Inventory (BDI-II) score using acoustic measurements in a sample of IIUM Engineering students. IOP Conference Series: Material Science Engineering [Internet] 2017 [cited 2019 May 19]; 260: 012022. Available from https://iopscience.iop.org/article/10.1088/1757-899X/260/1/012022/pdf
14. Yee NY, Sulaiman, WSW (2017) Resilience as mediator in the relationship between family functioning and depression among adolescents from single parent families. AKADEMIKA. 2017;87(1):111-22. http://doi.org/10.17576/akad-2017-8701-08.
15. Kottorp A. (2003). Occupation-based evaluation and intervention: validity of the assessment of motor and process skills when used with persons with mental retardation. [Unpublished PhD thesis]. Umee University, Sweden. 2003 [cited 2019 May 28]. Available from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A763461&dswid=1424
16. Sick J. Rasch measurement in language education Part 5: Assumptions and requirements of Rasch requirements of Rasch measurement. SHIKEN: JALT Testing & Evaluation SIG Newsletter. 2010;14(2):23-9.
17. Lerdal A, Kottorp A, Gay CL, Grov EK, Lee KA. (2014). Rasch analysis of the Beck Depression Inventory-II in stroke survivors: A cross-sectional study. Journal of Affective Disorders. 2014;158:48-52. doi: 10.1016/j.jad.2014.01.013.
18. Hammond SM. An IRT investigation of the validity of non-patient analogue research using the Beck Depression Inventory. European Journal of Psychological Assessment. 1995;11:14-20. doi: 10.1027/1015-5759.11.1.14.
19. Crocker L, Algina J. Introduction to classical and modern test theory. 1986. 527 p.
20. Thomas Rusch T, Lowry PB, Mair P, Treiblmaier H. Breaking free from the limitations of classical test theory: developing and measuring information systems scales using item response theory, Information & Management (I&M), 2017;54(2),189-203. doi:10.1016/j.im.2016.06.005.
21. Bond TG, Fox CM. Applying the Rasch Model: fundamental measurement in the human sciences (3rd ed.). 2015. 384 p.
22. Wright, BD. IRT in the 1990s: Which models work best? 3PL or Rasch? Rasch Measurement Transactions, 1992;6(1), 196-200.
23. de S6 Junior AR, de Andrade AG, Andrade LH, Gorenstein, Wang YP. Can gender and age impact on response pattern of depressive symptoms among college students? a differential item functioning analysis. Frontier in Psychiatry. 2019; 10: 50. doi: 10.3389/fpsyt.2019.00050
24. Khalaf ZF, Low WY, Merghati-Khoei E. Sexuality education in Malaysia: Perceived issues and barriers by professionals. Asia Pacific Journal of Public Health. 2014;26(4):358-66. doi: 10.1177/1010539513517258.
25. Linacre JM. Winsteps® Rasch measurement computerprogram. Beaverton. [Internet]. 2012 [cited 2019 Mac 10]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.359.6282&rep=rep1&type=pdf
26. Linacre JM. Investigating rating scale category utility. Journal of Outcome Measurement. 1999;3(2):103-22.
27. Linacre JM. A user's guide to Winsteps: Rasch model computer programs. Chicago: Winsteps. [Internet]. 2006 [cited 2019 Mac 1]. Available fromhttps://www.winsteps.com/winman/copyright.htm
28. Wang WC, Yao G, Tsai YJ, Wang JD, Hsieh C L. Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. Quality of Life Research. 2006;15(4):607-20. doi: 10.1007/s11136-005-4365-7.
29. Bilsen J. Suicide and youth: Risk factors. Frontiers in Psychiatry. 2018;9:540. doi/10.3389/fpsyt.2018.00540.
30. Latiff LA, Tajik E, Ibrahim N, Abubakar AS, Ali SSA. Depression and its associated factors among secondary school students in Malaysia. Southeast Asian Journal of Tropical Medicine & Public Health. 2016;47(1):131-41.
31. Wahab S, Rahman FNA, Hasan WMHW, Zamani IZ, Arbaiei NC, Khor SL, Nawi AM. Stressors in secondary boarding school students: Association with stress, anxiety and depressive symptoms. Asia-Pacific Psychiatry. 2013;5:82-89. doi: 10.1111/appy.12067.
32. Ministry of Education. Quickfact 2018: Malaysian education statistics. Putrajaya: Ministry of Education. [Internet] 2018 [cited 2019 May 19]. Available from https://www.moe.gov.my/index.

php/muat-turun/penerbitan-dan-jurnal/terbitan/buku-informasi/1587-quick-facts-2018-malaysia-educational-statistics-1/file
33. Wright BD, Linacre JM. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions. [Internet] 1994 [cited 2019 May 19];8:370-371. Available from https://www.rasch.org/rmt/rmt83b.htm .
34. Linacre JM. What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions [Internet] 2002 [cited 2019 May 19];6(2):878. Available from https://www.rasch.org/rmt/rmt162f.htm.
35. Hong S, Wong EC. Rasch rating scale modeling of the Korean version of the Beck Depression Inventory. Educational and Psychological Measurement. 2005;65(1):124-39. doi: 10.1177/0013164404267282.
36. Flaherty JA, Gaviria FM, Pathak D, Mitchell T. Developing instruments for cross-cultural psychiatric research. Journal of Nervous & Mental Disease. 1988;176:257-63.
37. Salokangas RK, Vaahtera K, Pacriev S, Sohlman B, Lehtinen V. Gender differences in depressive symptoms. An artefact caused by measurement instruments? Journal of Affect Disorder. 2002: 68(2-3):215-20.
38. Bulhxes C., Ramos E, Severo M, Dias S, Barros H. Measuring depressive symptoms during adolescence: What is the role of gender? Epidemiology and Psychiatric Sciences, 2019; 28(1):66-76. doi:10.1017/S2045796017000312.
39. de S6 Junior AR, de Andrade AG, Andrade LH, Gorenstein, Wang YP. Can gender and age impact on response pattern of depressive symptoms among college students? a differential item functioning analysis. Frontier in Psychiatry. 2019; 10: 50. doi: 10.3389/fpsyt.2019.00050.