Universal Journal of Educational Research 7(12A): 10-17, 2019 http://www.hrpub.org
DOI: 10.13189/ujer.2019.071902

Rasch Strategies for Evaluating Quality of the Conceptions and Alternative Assessment Survey (CETAS)

Rohaya Talib[1], Noorminshah A Iahad[2], Zakiah Mohamad Ashari[1], Mohd Rustam Mohd Rameli[1], Zainudin Abu Bakar[1,*], Rozilawati Dollah[2]

[1]School of Education, Universiti Teknologi Malaysia, Malaysia
[2]School of Computing, Universiti Teknologi Malaysia, Malaysia

Received November 6, 2019; Revised November 22, 2019; Accepted December 3, 2019

Abstract Due to society demand for educational development, education in Malaysia has begun to utilize alternative assessment approach in schools and universities. This study developed Conceptions and Alternative Assessment Survey (CETAS) to examine lecturers' conceptions of assessment (AC) and their practice of alternative assessment (AAP). In order for CETAS to be useful, a pilot study was conducted to examine quality of items in using Rasch Analysis approach. A total of 38 lecturers involved in this study. After item analysis, this study found that four items, Item 7 (AC), Item 8 (AC), Item 16 (AAP) and Item 30 (AAP) did not meet the requirement of fit statistics analysis and local item dependency. Therefore, four items were deleted while other 58 are suitable to be used for measuring the intended constructs. In addition, the scale calibration analysis also revealed that Scale 3 (slightly disagree) was not well-functioning. Therefore, after consideration of analysis and expert review, Scale 3 was collapsed leaving CETAS with 5 scales. Nevertheless, CETAS has a good item and person reliability and can be used to examine lecturers' conceptions of assessment and their practices of alternative assessment.

Keywords Alternative Assessment, Assessment, Rasch Analysis

1. Introduction

Recently, alternative assessment has regained increasing attention after it was first introduced in 1990 (Quenemoen, 2008). In general, assessment is defined as a measure of performance including knowledge, skills, attitudes and beliefs. Generally, for an educator, the major purposes of assessment merely revolve around classroom such as diagnosing students' strength and weakness, monitoring students' progress, assigning students' grades and determining their own instructional effectiveness (Popham, 2014; Green & Johnson, 2010). In addition, Popham (2014) listed three more important purposes of assessment such as influencing public perceptions of educational effectiveness, evaluating educators, and clarifying instructional intentions.

Traditional assessment which usually employed pen and paper tests has few limitations; it can only measure what learners can do at one particular time (Law & Eckes, 1995; Dikli, 2003); focus on lower level thinking skills such as knowledge and comprehension (Dikli, 2003; Ince & Yilmaz, 2012); no immediate feedback is given to the learners (Bailey & Brown, 1998); the purpose of traditional assessment is usually norm-referenced (Dikli, 2003) and it does not necessarily reflect students' own experience (Brempong, 2019). Generally, traditional assessments are for assigning students grade, diagnosing students' strength and weakness. In contrast, alternative assessment is also known as non-formal testing and focuses on assessing higher-order thinking skills (Bagley, 2010), life-long skills (Ince and Yilmaz, 2012), real-life tasks (Quansah, 2018) which allow students to demonstrate best what they have learned (Quansah, 2018). In addition, alternative assessment allows instructors to evaluate what students can do and cannot do instead of what they know and do not know. Alternative assessment also has been evidently reported to motivate students to learn (Tal & Miedijensky, 2005; Bachelor, 2015). In general, alternative assessments are useful for monitoring students' progress and determining instructional effectiveness.

In Malaysia, a shift of assessment approaches from traditional assessment to alternative assessment occurs in both schools and universities. This is due to society demand for educational development in order to move towards a more powerful learning environment. Assessment has always been a responsibility of the instructors. Watkins et al. (2005) pointed out instructors' conception of assessment is influenced by their viewed of theories of teaching and learning. Additionally, Badariah et al., (2014) study revealed that lecturers in higher education have limited practice of assessment for learning. This may be due to unfamiliarity with formative assessment. The study did not further examine lecturers' view on conceptions and alternative assessment. Therefore, this study developed a survey to examine lecturers' perspective on conceptions of alternative assessment and focus on assessing quality of items in CETAS using Rasch Measurement Analysis (Rasch, 1960).

2. Rasch Measurement Model

Rasch Measurement Model (RMM) is item response theory-based model that has been used numerously in evaluating item quality. RMM estimates only one parameter; difficulty parameter. While item discrimination and guessing are assumed to be constant (Magno, 2009). One of RMM strengths is its measurement requirement in which items in any instrument should fit the model well enough to produce useful measures (Boone & Noltemeyer, 2017). Another strength of RMM is that it can convert nonlinear raw data to linear scale once the requirement is met (Boone & Noltemeyer, 2017; Boone, 2016). Therefore, once the requirement is met, researchers can make meaningful interpretation of their survey score.

There are varieties of ways of how RMM can be used to evaluate items quality of social science research instrument (Bond, 2003; Boone & Noltemeyer, 2017; Boone, 2016). Among analyses that RMM offers are; (i) examining internal consistency, (ii) examining fit analysis statistics, (iii) examining unidimensionality and item dependency, and (iv) scale calibration. Internal consistency includes person and item reliability and separation. Fit analyses include examining point measure correlation, mean-square error (MNSQ), and standardized fit statistic (ZSTD). As for unidimensionality, the analysis should focus on its eigenvalue. Lastly, to measure local item dependency, Rasch Analysis can examine residual correlation. As for the analysis guideline, Table 1 shows requirements for all analyses above.

Table 1. Analysis Requirement

Analysis	Requirement	Interpretation
Item Reliability and Separation	Reliability should be larger than 0.80 Separation index should be larger than 2.00	Item reliability of 0.80 with separation more than 2.00 indicates that the sample is able to confirm the item hierarchy to at least two levels.
Person Reliability and Separation		Person reliability of 0.80 with separation more than 2.00 indicates items are sensitive enough to separate at least two levels of person ability.
Point-measure Correlation	Value should be positive, at least more than 2.0	Negative value indicates that the item contradicts the direction of the latent trait
MNSQ	0.5 to 1.50	Productive for measurement
ZSTD	-2.00 to 2.00	Reasonable predictability
Eigenvalue	Should not be larger than 3	Eigenvalue larger than 3 indicates some kind of secondary effect
Residual Correlation	Should not be larger than 0.70	The correlation that is more than 0.70 indicates the item is independent with another item.

3. Methodology

3.1. Instrument

Conceptions of Alternative Assessment Survey (CETAS) have two main dimensions; conception on assessment in general and also alternative assessment practice. For assessment conceptions; CETAS measures how lecturers view assessment in terms of improvement on teaching and learning, institutional accountability, irrelevances, and student accountability, while for alternative assessment practice; CETAS also measures lecturers' alternative assessment practice approach such as authentic assessment, challenge-based, integrated, performance, personalized, profiling, project-based and real-time. The distribution of items is shown in **Table 2.**

Table 2. Tabulation of Items

Dimension	Sub-dimension	Items	Total
Conception	Improvement on Teaching and Learning	1,2,9,10	4
	Students Accountability	5,6,14,15,16	5
	Institutional Accountability	3,4,11,12,13	5
	Irrelevance	7,8,17,18	4
Practice	Authentic Assessment	1,9,17,25,33	5
	Challenge-based	7,15,23,31,39,43	6
	Integrated	4,12,20,28,36	5
	Performance	3,11,19,27,35	5
	Personalized	2,10,18,26,34,41	6
	Project-based	5,13,21,29,37,42	6
	Profiling	8,16,24,32,40,44	6
	Real-time	6,14,22,30,38	5
Total Items			62

3.2. Sample

The pilot study involved 38 lecturers from Universiti Teknologi Malaysia. The number of respondent required by Rasch Analysis is 30 persons. According to Wright and Tennant (1996), 30 well-targeted samples are enough to evaluate items quality.

4. Results

4.1. Internal Consistency

As shown in **Table 3**, overall analysis indicates that Conceptions of Alternative Assessment Survey (CETAS) has a person reliability of 2.67 and person separation of 0.88. Based on the person separation and person reliability, items in CETAS are able to separate people to 2.67 ≈ 3 level. Person separation that is more than 2 and person reliability that is more than 0.80 indicate that CETAS is sensitive enough to distinguish between low and high respondents.

As for item analysis, item separation indicates that how able person sample is to distinguish items between different levels of difficulty. CETAS has item reliability of 0.94 and separation of 3.90. Item separation that is more than 2 and person reliability that is more than 0.80 imply that the person sample is large enough to confirm the hierarchy of the items in CETAS. In this pilot study, the total of respondents is 38, which means that 38 respondents are enough to confirm the hierarchy of the items in CETAS.

Table 3. Summary Statistics

Summary Statistics	Value
Person Separation	2.67
Person Reliability	0.88
Item Separation	3.90
Item Reliability	0.94

4.2. Unidimensionality

To examine unidimensionality, Linacre (2015) suggested to examine Unexplained Variance in the 1st Contrast. Eigenvalue should not be larger than 3. However, as shown in **Table 4**, CETAS has an eigenvalue of 9.0 which is two times larger than required value. This indicates that CETAS measures multidimensionality. It shows that CETAS did not measure only one dimension. The result is reasonable as CETAS has two main dimensions which are Assessment Conception (AC) and Alternative Assessment Practice (AAP). In AC, there are four sub-dimensions such as 'Improvement on Teaching and Learning', 'Institutional Accountability', 'Irrelevance' and 'Student Accountability'. As for AAP, there are eight sub-dimensions which measure 'Authentic', 'Challenge-based', 'Integrated', 'Performance', 'Personalized', 'Profiling', 'Project-based' and 'Real-time'. Therefore, this study further examines each of these sub-dimensions.

Table 4. Standardized Residual Variance

Indices	Empirical	Modeled
Raw variance explained by measures	46.1%	47.1%
Indices	**Eigenvalue**	**Percentage**
Unexplained variance in 1st contrast	9.0	7.8%

4.3. Unidimensionality of Assessment Conception

Further analysis indicates that all eigenvalue of the Assessment Conception sub-dimensions is less than 3 as shown in **Table 5**. However, the percentage of raw unexplained variance each sub-dimension has exceeded 15% limit. Therefore, further analysis of local item dependency should be conducted. The correlation between a pair of items should not exceed 0.70. If this correlation occurred, one of the items should be removed as both of the items probably shared the same dimension.

Table 5. Standardized Residual Variance (Assessment Conception)

Construct	Raw Variance Explained by Measure		Raw unexplained variance (1^{st} Contrast)	
	Empirical	**Modeled**	**Eigenvalue**	**Percentage**
Improvement on Teaching and Learning	58.1%	58.7%	2.1	22.2%
Student Accountability	40.9%	40.1%	2.0	23.5%
Institutional Accountability	45.9%	46.5%	1.9	20%
Irrelevance	77.1%	74.0%	2.0	11.6%

To identify dependent item, Winstep produces the largest standardized residual correlation table between a pair of items. **Table 6** shows correlation between pairs of Assessment Conception items that have large correlation. As can be seen in **Table 6**, correlation between Item 7 and Item 18 is quite large with a correlation of 0.69. The correlation between Item 1 and Item 3 is also quite large as well with a correlation of 0.68. Though all correlation is less than 0.70, Item 7, Item 8, and Item 18 have appeared more than two times in the **Table 6**. Therefore, Item 7, Item 8 and Item 18 are flagged as possible items to be omitted.

Table 6. Largest Standardized Residual Correlation

Items Pair		Correlation
Item 7	Item 18	0.69
Item 1	Item 3	0.68
Item 7	Item 8	0.63
Item 2	Item 3	0.63
Item 8	Item 18	0.55
Item 1	Item 2	0.54
Item 14	Item 15	0.53
Item 8	Item 11	-0.64
Item 10	Item 18	-0.63
Item 7	Item 11	-0.58

Further analysis examines fit statistics of each Conception Assessment items. Fit statistics include the examination of point measure correlation, mean-squared error, and standardized fit statistics. Point measure correlation should be more than 0.20. Table 1.5 shows Item 8 has a negative point measure correlation. **Table 7** also shows that Item 7 has an unfit value for all fit indices. Point measure correlation value of Item 7 is also small with a value of 0.02. Additionally, Item 7 also appeared in local item dependency table. After considering all fit indices and local item dependency, Item 7 will be omitted from Conception Assessment items.

Table 7. Fit Statistics Analysis (Conception Assessment)

Indices	Range	Misfit Item
Point Measure Correlation	0.58 to -0.02	Item 8 (-0.02), Item 7 (0.03)
MNSQ Outfit	0.43 to 4.10	Item 7 (4.10)
ZSTD Outfit	-2.60 to 6.90	Item 7 (6.90), Item 8 (2.90), Item 6 (-2.60)
MNSQ Infit	0.45 to 2.77	Item 7 (2.77)
ZSTD Infit	-2.60 to 5.20	Item 7 (5.20), Item 8 (2.20), Item 6 (-2.60)

4.4. Unidimensionality of Alternative Assessment Practice

The second dimension is Alternative Assessment Practice. As can be seen in **Table 8**, eigenvalue of each sub-dimension is less than 3. However, the percentage of the eigenvalue is more than the limit 15%. Therefore, local dependency of Alternative Assessment Practice item is examined.

Table 8. Standardized Residual Variance (Alternative Assessment)

Construct	Raw Variance Explained by Measure		Raw unexplained variance (1^{st} Contrast)	
	Empirical	Modeled	Eigenvalue	Percentage
Authentic	38.7%	38.7%	1.9	23.7%
Challenge-based	41.6%	39.3%	1.9	18.3%
Integrated	39.7%	41.2%	1.4	17.4%
Performance	44.8%	44.2%	1.9	20.7%
Personalized	55.5%	53.6%	2.3	16.9%
Profiling	38.1%	38.7%	2.6	26.7%
Project-based	45.3%	44.6%	1.7	15.4%
Real-time	41.7%	40.7%	1.6	18.4%

Local item dependency as shown in **Table 9** shows that two pairs of item have correlation more than 0.70, which is Item 8 and Item 16, and Item 8 and Item 9. Item 8 obviously has a high correlation between two items, which is Item 16 and Item 9. Thus, Item 8 is considered locally dependent. Therefore, Item 8 will be omitted.

Table 9. Largest Standardized Residual Correlation

Items Pair		Correlation
Item 8	Item 16	0.76
Item 8	Item 9	0.75
Item 26	Item 34	0.59
Item 9	Item 16	0.56
Item 14	Item 16	0.55
Item 8	Item 11	0.54
Item 8	Item 15	-0.58
Item 15	Item 25	-0.58
Item 14	Item 34	-0.57
Item 3	Item 13	-0.55

Fit analysis statistics are further conducted to examine any other items that are unfit to measurement. **Table 10** shows fit analysis for alternative assessment practice items. Based on the table, only two items did not fulfil all criteria of fit statistics which are Item 16 and Item 30. Therefore, Item 16 and Item 30 are considered unfit thus they will be omitted from this questionnaire.

4.5. Rating Scale Calibration

Rasch Analysis also allows researcher to calibrate instrument scale. It provides information whether one of the scales should be collapsed or another scale should be included. Conceptions and Alternative Assessment Survey (CETAS) have six scales which are:

1. Strongly Disagree
2. Disagree
3. Slightly Disagree
4. Slightly Agree
5. Agree
6. Strongly Agree

There are few ways to examine the rating scale as shown in **Table 11**. Firstly, it's by examining observed count and observed average. As for observed count, a minimum of 10 observations if required for observed count in each scale with a fair distribution across the rating scales (Linacre, 2012). As for observed average, the indice should steadily and consistently increase. Secondly, by examining structure calibration. The structure calibration is manually calculated between two scales. The difference between two scales should not be less than 1.4 and should not be more than 5 (Linacre, 2012). A value below 1.4 indicates overlapping between categories and the respondents are unable to differentiate the scales. Lastly is by observing probability curve pattern. Each scale should have a distinct peak to indicate the rating scale is well-functioning.

Table 10. Fit Statistics Analysis (Alternative Assessment Practice)

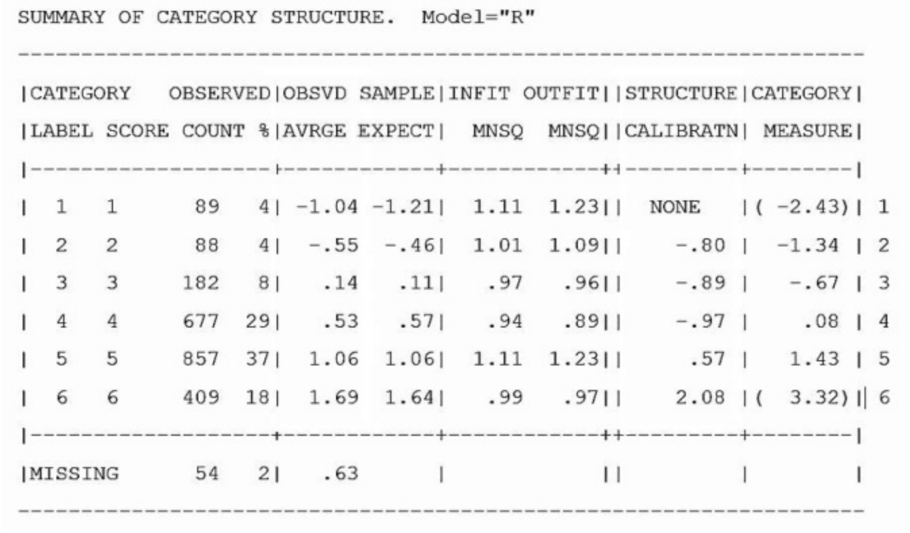
Indices	Range	Acceptable Value	Misfit Item
Point Measure Correlation	0.72 to 0.13	Should be larger than 0.20	Item 34 (0.13), Item 16 (0.15), Item 39 (0.19), Item 30 (0.19)
MNSQ Outfit	0.41 to 1.73	0.50 to 1.50	Item 39 (1.73), Item 26 (1.72), Item 34 (1.71), Item 44 (1.70), Item 16 (1.70), Item 30 (1.63), Item 10 (0.41)
ZSTD Outfit	-2.80 to 2.90	-2.00 to +2.00	Item 39 (2.80), Item 26 (2.90), Item 34 (2.30), Item 44 (2.30), Item 16 (2.30), Item 30 (2.20), Item 10 (-2.80)
MNSQ Infit	0.43 to 1.67	0.50 to 1.50	Item 34 (1.60), Item 44 (1.63), Item 16 (1.67), Item 30 (1.65), Item 10 (0.43)
ZSTD Infit	-2.70 to 2.20	-2.00 to +2.00	Item 44 (2.10), Item 16 (2.20), Item 30 (2.10), Item 1 (-2.10), Item 10 (-2.70)

Table 11. Indicators of a rating scale

Indicators	Descriptions of a rating scale
Observed count	High and stable observed count. Low values often indicate unnecessary or redundant categories
Observed average	Expected to increase in size as the category increases
Structure calibration	Expected to increase in size as the category increases Expected difference between threshold is 1.4 < x < 5
Probability curve	Each category is expected to have distinct peak

As can be seen in **Figure 1**, observed average of first scale is -1.04 and the average steadily increases up to 1.69 at the last scale. However, observed count for scale 2 is decreased. The observed counts for scale 1 and 2 are lowest among all of the counts.

Table 12 shows structure calibration difference between scales. There are calibration differences that are out of the required range which are between scale 2 and scale 3 and also between scale 3 and scale 4. Additionally, **Figure 2** also shows that scale 2 is overshadowed by scale 1, while scale 3 is overshadowed by category 4. Both scale 2 and scale 3 have no distinct peak compared to other scales. This indicates that scale 2 and scale 3 are not well functioning. The respondents were unable to differentiate these scales.

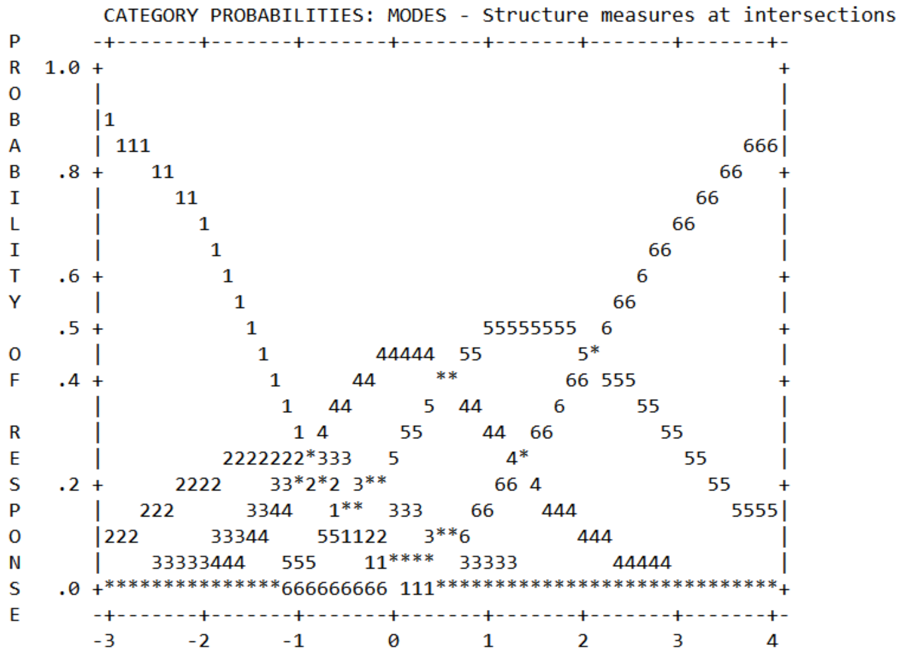


```
SUMMARY OF CATEGORY STRUCTURE.  Model="R"
-------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN| MEASURE|
|-------------------+------------+------------++---------+--------|
|  1   1      89   4| -1.04 -1.21|  1.11  1.23||  NONE   |( -2.43)| 1
|  2   2      88   4|  -.55  -.46|  1.01  1.09||    -.80 |  -1.34 | 2
|  3   3     182   8|   .14   .11|   .97   .96||    -.89 |   -.67 | 3
|  4   4     677  29|   .53   .57|   .94   .89||    -.97 |    .08 | 4
|  5   5     857  37|  1.06  1.06|  1.11  1.23||     .57 |   1.43 | 5
|  6   6     409  18|  1.69  1.64|   .99   .97||    2.08 |(  3.32)| 6
|-------------------+------------+------------++---------+--------|
|MISSING      54   2|   .63      |            ||         |        |
-------------------------------------------------------------------
```

Figure 1. Summary of Category Structure

Table 12. Calibration Differences

Scale	Structure Calibration	Calibration Differences
1	-	-
2/3	-0.89-(-0.80)	-0.09
3/4	-0.97-(-0.89)	-0.08
4	0.57-(-0.97)	1.54
5	2.08-0.57	1.51



```
        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      -+-------+-------+-------+-------+-------+-------+-------+-
R  1.0 +                                                         +
O      |                                                         |
B      |1                                                        |
A      | 111                                                  666|
B   .8 +    11                                              66   +
I      |      11                                          66     |
L      |        1                                       66       |
I      |         1                                    66         |
T   .6 +          1                                  6           +
Y      |           1                               66            |
    .5 +            1                   55555555  6              +
O      |             1         44444  55        5*               |
F   .4 +              1      44     **         66 555            +
       |               1   44      5  44      6      55          |
R      |                1 4      55     44  66         55        |
E      |          2222222*333   5         4*             55      |
S   .2 +      2222    33*2*2 3**         66 4              55    +
P      |   222      3344   1**  333    66    444             5555|
O      |222      33344    551122   3**6         444              |
N      |    33333444   555    11****  33333        44444         |
S   .0 +***************666666666 111*****************************+
E      -+-------+-------+-------+-------+-------+-------+-------+-
       -3      -2      -1       0       1       2       3       4
```

Figure 2. Category Probability

5. Discussion

This study developed an instrument to measure lecturers' conception towards assessment and also their alternative assessment practice. Conceptions of Alternative Assessment (CETAS) has been responded by 38 lecturers from Universiti Teknologi Malaysia (UTM). In order for CETAS being useful for operational use, item analysis was conducted to evaluate quality of item using Rasch Analysis. There were four item analyses conducted; (i) examining internal consistency, (ii) examining unidimensionality and local item dependency, (iii) examining fit statistics and (iv) scale calibration.

Based on the value of person and item reliability and separation, items in CETAS are sensitive enough to separate respondents into three levels of ability. One of the characteristics of good items is it can discriminate respondents into at least two different abilities. Unidimensionality and local item dependency complement each other. Violation of local item dependency may affect unidimensionality of an instrument. As for CETAS, overall unidimensionality analysis indicated that there was a sign of multidimensionality. However, it does make sense as CETAS has two different dimensions; conceptions (AC) and practice (AAP). Therefore, unidimensionality analysis was conducted separately for assessment conception and alternative assessment practice. Any items that were deemed inappropriate are deleted to preserve unidimensionality of CETAS. Any items that did not fulfil requirement for local item dependency and fit statistics were deleted. Therefore, total four items were deleted, Item 7 (AC), Item 8 (AC), Item 16 (AAP) and Item 30 (AAP).

Lastly, based on scale calibration analysis, respondents could not differentiate between scale 3 and scale 4 (disagree). In addition, scale 3 also has no peak, which means the scale is not well functioning. Therefore, based on analysis and expert review on these scales, scale 3 (slightly disagree) was collapsed. For operational use in the future, CETAS will only have five scales. After deletion, CETAS can be used in the future to examine lecturers' conceptions of assessment and their practice of alternative assessment.

REFERENCES

[1] Bachelor, R. B. (2015). Alternative Assessments and Student Perceptions in the Foreign Language Classroom. Ed.D. Dissertations. Olivet Nazarene University.

[2] Badariah, T., Zubairi, A. M., Ibrahim, M. B., Othman, J., Rahman, N. S. A., Rahman, Z. A., ... & Ahmad, T. (2014). Assessment for learning practices and competency among Malaysian university lecturers: A national study. Practitioner Research in Higher Education, 8(1), 14-31.

[3] Bagley, S. S. (2010). Students, teachers and alternative assessment in secondary school: Relational models theory (RMT) in the field of education. The Australian Educational Researcher, 37(1), 83-106.

[4] Bailey, K., & Brown, J. (1999). Learning about language assesment: Dilemmas, decisions, and directions & new ways of classroom assessment. Learning, 4(2), 1-8.

[5] Brempong, D.A. (2019). Comparing Traditional Assessment Procedures, and Performance and Portfolio Assessment Procedures. DOI: 10.13140/RG.2.2.20943.12960

[6] Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how?. CBE—Life Sciences

Education, 15(4), rm4.

[7] Boone, W. J., and Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. Cogent Education, 4(1), 1416898.

[8] Bond, T. G. (2003). Relationships between cognitive development and school achievement: A Rasch measurement approach. On the forefront of educational psychology, 37-46.

[9] Dikli, S. (2003). Assessment at a Distance: Traditional vs. Alternative Assessments. *Turkish Online Journal of Educational Technology-TOJET*, *2*(3), 13-19.

[10] Green, S.K and Johnson R.L. (2010). Assessment Is Essential. (1st ed). New York: McGraw-Hill

[11] Ince, E., & Yilmaz, O. (2012). The Usage of Alternative Assessment Techniques in Determination of Misconceptions About Electromagnetic Field-Magnetism Contents and Effects of Video-Based Experiments on Pre-service's Achievement. Procedia-Social and Behavioral Sciences, 55, 206-211.

[12] Law, B., & Eckes, M. (1995). Assessment and ESL: On the Yellow Big Road to the Withered of Oz. A Handbook for K-12 Teachers. Peguis Publishers Limited, 100-318 McDermot Avenue, Winnipeg, Manitoba, Canada R3A 0A2.

[13] Linacre, J. M. (2012). Winsteps help for Rasch analysis. Retrieved from Winsteps website: http://www. winsteps. com.

[14] Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. The International Journal of Educational and Psychological Assessment, 1(1), 1-11.

[15] Popham, W.J. (2014). Classroom Assessment What Teachers Need to Know. (7th ed). United States of America: Pearson.

[16] Quansah, F. (2018). Traditional or Performance Assessment: What is the Right Way in Assessing Learners? Research on Humanities and Social Sciences. Vol.8, No.1, 2018 21

[17] Quenemoen, R. (2008). A Brief History of Alternate Assessments Based on Alternate Achievement Standards. Synthesis Report 68. *National Center on Educational Outcomes, University of Minnesota*.

[18] Tal, R. T., & Miedijensky, S. (2005). A model of alternative embedded assessment in a pull-out enrichment program for the gifted. Gifted Education International, 20(2), 166-186.

[19] Watkins, D., Dahlin, B., & Ekholm, M. (2005). Awareness of the backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lecturers. Instructional Science, 33(4), 283-309.