A Validation of the Assessment Practices Inventory Modified (APIM) Scale using Rasch Measurement Analysis

Matovu Musa
Islamic University in Uganda
Senior Lecturer, Faculty of Education
Email: matovumousa@yahoo.com; m.matovu@iuiu.ac.ug

Abstract

There are many instruments that have been designed to measure assessment practices skills, but very few have been validated for their soundness and consistency in measuring lecturers' assessment practices skills. This study was undertaken to examine the psychometric properties of the Assessment Practices Inventory Modified (APIM) scale, and its soundness in measuring assessment practices skills among university lecturers. A quantitative survey research design was adopted for this study. The 50-item APIM scale on a five-point Likert scale was administered to a sample of 321 lecturers randomly selected from six universities in Uganda. The data collected was analysed using WINSTEPS Rasch Measurement Modelling Program for both Classical Test Theory (CTT) and Item Response Theory (IRT) to test the psychometric properties of the APIM scale. From the results of both the CTT (Cronbach's alpha and the point bi-serial coefficients) and IRT (category probability curve, item and persons' reliabilities, item characteristic curve, item difficulty, fit statistics, and principal component analysis) in this study, the APIM scale was found to have adequate psychometric properties in measuring assessment practices skills among university lecturers. The APIM scale was also found to be invariant to gender of the university lecturers. In conclusion, the APIM scale has been found to be sound and consistent in measuring university lecturers' assessment practices skills. This study has pronounced a sound and consistent instrument in measuring assessment practices skills among university lecturers in Uganda, and has provided universities in Uganda with a valid and reliable instrument which will measure assessment practices skills of their lecturers. The results of this study have highlighted that the APIM scale can universally measure assessment practices skills among university lecturers.

Keywords: validation, assessment practices inventory modified, Rasch measurement analysis

Assessment is an interaction in which lecturers closely observe and collect information about student learning, interpret it, and apply the results from the assessments (Fatmawati, 2011; Singh, Lebar, Kepol, Abdul-Rahman, & Mukhtar, 2017; Orzolek, 2006). In the assessment process, lecturers gather what students have learnt in order to measure whether the learning goals have been achieved (MOE, 2010; NCCA, 2005; Satterly, 1989). The information obtained through assessments helps lecturers to understand the students' levels of learning, helps in making academic decisions, and can be used to improve

student learning (ACRL, 2017; Ainsworth & Viegut, 2006; Hassel & Ridout, 2018; Hindi & Miller, 2000; Martell & Calderon, 2005; Popper, 2005). The importance of assessments in student learning necessitates having an instrument with high precision to measure the lecturers' skills in assessing students. This would help in controlling quality in student learning.

Assessment practices skills are proficiencies in which lecturers design tests, award scores/ grades, analyse them, and use the results from the assessments to improve student learning (Ainsworth & Viegut, 2006; Brown, 2003; Coates, 2015; McClarty & Gaertner, 2015). The Assessment Practices Inventory Modified (APIM) scale validated in this study measures the lecturers' assessment design, administration, interpretation, and application skills in assessing students. According to the different studies done on assessment practices, there is lack of evidence on validated instruments used to collect information about lecturers' assessment practices skills (Connoley, 2004; Greatorex, Johnson, & Coleman, 2017; Lai et al., 2015; O'Donovan, Price, & Rust, 2001; O'Grady, 2006; Price & Rust, 1999; Rousselot et al., 2018; Russell & Markle, 2017; Watkins, 1998). Therefore, it is important to get an appropriate instrument that can measure university lecturers' assessment practices skills adequately. The purpose of this study was to validate whether the APIM scale is adequate in measuring assessment practices skills among university lecturers. The validation was to understand whether the categories and items in the APIM scale were functioning well and contributing to the uni-dimensionality of the instrument. The study also examined whether the APIM scale is invariant to gender as a moderating variable.

Literature Review

The quality of results from assessments in universities can only be guaranteed to be accurate and reflecting the learning objectives if there is an appropriate instrument to measure the lecturers' skills in the assessing students. Several studies have been conducted to examine university lecturers' skills in assessing students (Alkharusi, 2012; Biggs, 2003; McMillan, 2001), but no study has clearly defined an instrument that can measure university lecturers' assessment practices skills adequately. According to the literature of assessment practices, different theories have been developed to explain how university lecturers can assess their students. Different theories highlight different skills in assessment practices, but all the skills in the various theories can be summed up into four major components; design, administration, interpretation and application (Ainsworth & Viegut, 2006; Braney, 2010; Burry-Stock & Frazier, 2008).

From the various modern theories that explain assessment practices, a theory proposed by Ainsworth and Viegut (2006) can be used to explain the skills of design, interpretation, and application in assessment practices (Ainsworth & Viegut, 2006; Braney, 2010). The literature of institutional support for student assessment explains the component of administration in

student assessment (Peterson, Einarson, Augustine, & Vaughan, 1999a). Ainsworth and Viegut (2006) claim that the student assessment process comprises of the way an assessment is designed, interpreted, and how the assessment results are used or applied. Peterson et al. (1999a) highlight on how administration is an important aspect in the student assessment process. In the previous studies done on assessment practices, it is mentioned that the Ainsworth and Viegut (2006) theory explains student assessment better than any other modern assessment theory (Braney, 2010).

Design

Design is the first level of the assessment process which prepares ground for other levels in this process. Designing of an assessment should be done in contemplation of how the assessment will be administered, scored and interpreted, and how the different stakeholders will use the feedback. A well-designed assessment should be set according to the priority standards, have unwrapped standards, and multiple measures (Ainsworth, 2003a; Ainsworth & Viegut, 2006). *Priority standards* are the essential components students should know and be able to execute in an assessment while *unwrapped standard*s are concepts and skills lecturers use to measure the students' progress towards a particular standard (Ainsworth, 2003b). *Multiple measures* is the use of different types of questions in an assessment which helps to capture an accurate picture of the student learning (Braney, 2010; Marzano, 2003).

Interpretation

During the interpretation of assessment, lecturers would like to understand the learning outcomes of students as individuals and as groups in a particular course. Interpretation of assessments involves scoring and analysing of the assessment results. In scoring an assessment, students are awarded scores/ grades according to their academic efforts while analysing is the organising of assessment results in order to draw meaning out of them. Scoring and interpreting of assessment results requires assessment expertise to avoid misinterpretations in the students' scores or grades (O'Connor, 2009). Interpretation of assessments can be undertaken by individual lecturers or as a group in order to make conclusive decisions about student performance (Reeves, 2004). Analysing results as a group builds consensus with regard to students' proficiency (Braney, 2010).

Application

After assessment results have been obtained, they can be used for various purposes in the learning process. The obtained assessment results can be used to make education decisions, refine the next test for proficiency, design policy, and to improve the learning process (Ainsworth & Viegut, 2006; Peterson et al., 1999a). In application or use of student assessment results, this might also involve giving feedback to students on what they have obtained in the previous assessment (Reeves, 2004).

Administration

This is the management support given by the individual lecturers or by the administration of the university towards the assessment process (Peterson, Einarson, Augustine, & Vaughan, 1999b). The administrative support by the individual lecturers and institutions helps in the smooth running of the designing, scoring, and grading processes in the student assessment (Gibbs, 2006). The administrative support activities in the assessment process range from communicating results to students to controlling external influence in the assessment process.

Research Objectives

The general objective of this study was to examine the psychometric properties of the Assessment Practices Inventory Modified (APIM) scale. In the validation scale, the study specific objectives were;

1. To find out the reliability and validity of the APIM scale;
2. To examine the category and item functioning in the APIM scale; and
3. To examine whether the APIM scale is invariant to gender of the university lecturers.

Methods

This section highlights the techniques and methods used in conducting the validation of the APIM scale.

Measures

The APIM scale is an instrument designed to measure assessment practices skills among university lecturers in Uganda. To test the psychometric properties of the APIM scale, Rasch measurement model using WINSTEPS version 3.64.2 was used. Rasch measurement analysis using Item Response Theory (IRT) was used to test the APIM scale's soundness and consistency in measuring assessment practices skills, though it also gave results for Classical Test Theory (CTT). Rasch measurement analysis' ability to transform ordinal scores into logits and to measure the person's relative assessment skills onto the item difficulty on the same continuum made it the most appropriate method to validate the APIM scale (see Ainol-Madziah & Noor-Lide, 2006; Hambleton et al., 1991; Rasch, 1960; Rasch, 1966; Wright & Masters, 1982; Wright & Mok, 2000). Rasch measurement analysis was used to examine the Cronbach's alpha and the point bi-serial coefficients using CTT while the person and item reliabilities, fit statistics, item difficulty, person-item map, category function, and Differential Item Functioning (DIF) were measured using IRT to determine the APIM scale's precision in measuring university lecturers' assessment practices skills.

Design

This study involved three major phases; (a) developing the APIM scale, (b) data collection, and (c) conducting CTT and IRT analyses using Rasch measurement analysis. After developing the APIM scale it was given to 6 experts in the area of assessment and evaluation for content validation. The experts highlighted that the APIM scale was adequate to measure the lecturers' skills in assessment practices. Later, the APIM scale was administered to university lecturers in Uganda who participated in the study willingly and also, completed the scale anonymously.

Sample

A sample of 321 lecturers was selected using simple random sampling technique from six universities in Uganda. The universities were also randomly selected equally from both private and public universities, that is, three universities each from category. The sample was selected by putting into consideration of the the lecturers' gender as a moderating variable.

Instrumentation

The APIM scale which collected data for this study consists of 50 items (see appendix). The items in the scale were adopted and adapted from the Assessment Practices Inventory Revised scale (API_R) (Burry-Stock & Fraizer, 2008), the Assessment Practices Inventory scale (API) (Zhang & Burry-Stock, 1994), and from literature of Institutional Support for Student Assessment (Peterson et al., 1999a). The API_R and API were previously used to collect information about assessment practices skills of school teachers. The items of the APIM scale are divided into four subscales and are on a 5-point Likert-scale on which university lecturers rated the way they perceived their skills in terms of assessment practices. The APIM rating scale ranges from not at all skilled to highly skilled (1 = *Not-at-all-skilled*, 2 = A-*little-skilled*, *3 = Some-what-skilled*, *4 =Skilled*, and 5 = *Highly-skilled*). The underlying dimensions in the APIM include design (Items 6, 7, 17, 20, 23, 24, 29, 32, 33, 34, 35, 36, 37, and 38), administration (Items 1, 9, 14, 15, 18, 19, 22, 25, 28, 31, 41, 42, and 44), interpretation (Items 2, 3, 5, 13, 16, 21, 26, 27, 43, 45, 46, and 47), and application (Items 4, 8, 10, 11, 12, 30, 39, 40, 48, 49, and 50). The subscales in the APIM scale correspond to variables which have been mentioned in other literatures about assessment practices skills. The underlying dimensions of the APIM scale were extracted based on the results of Exploratory Factor Analysis when the 50 items in the scale were run in one model (see Table 1).

Table 1
Dimensions of the APIM Scale

Dimension	Description	Items
Design	Developing of assessments	6, 7, 17, 20, 23, 24, 29, 32, 33, 34, 35, 36, 37, 38
Administration	Management of assessments	1, 9, 14, 15, 18, 19, 22, 25, 28, 31, 41, 42, 44
Interpretation	Understanding assessment results	2, 3, 5, 13, 16, 21, 26, 27, 43, 45, 46, 47
Application	Use of assessment results	4, 8, 10, 11, 12, 30, 39, 40, 48, 49, 50

Validity and Reliability

Validity is referred to as the relevancy of the collected information to the issue being investigated while reliability is the stability or consistency of an instrument that undertakes a measurement (Airasian & Russell, 2008; Messick, 1989). Validity and reliability need to be clearly scrutinised in order to pronounce the precision of any measurement instrument. Conducting validity and reliability for the APIM scale was necessary because the scale was newly-developed. Rasch measurement analysis was used to examine the validity and reliability of the APIM scale. In this study, the reliability of the APIM scale was .93 which is very good. This indicated that the APIM scale had good precision in measuring skills in assessment practices of university lecturers.

Findings

In validating the APIM scale, the Category Probability Curve (CPC) was examined to understand how the items functioned in the scale. Examining the category functioning of a scale is done by predicting item values to the person locations on the latent trait (Andrich, 2005; Hagquist & Andrich, 2004a, b; Hagquist, 2001, 2008). The CPC was inspected to examine whether the APIM rating scale had well-defined, ordered, and simple categories to explain the university lecturers' assessment practices skills (see Linacre, 1999a, b; Shaw et al., 1992; Smith et al., 2003). Using Rasch measurement model the 50 items on 5 categories were examined on their category fit. The step values generated in the structure measure increased monotonically but did not show distinctiveness between categories 2-3 (-.67) and 3-4 (.20). This illustrated that the categories did not function as expected (see Bond & Fox, 2007; Linacre, 1999b, 2011) (see Table 2).

Table 2
Summary of the Category Structure of the 50 Items

Category label	Observed Counts	%	Average Measures	Infit MnSq	Outfit MnSq	Structure Measure	Category Measure
1	444	3	-.12	1.03	1.05	None	(-2.94)
2	2091	13	-.06	1.03	1.03	-1.61	-1.33
3	4900	31	.31	1.00	1.02	-.67	-.14
4	6507	41	.67	1.00	1.00	.20	1.28
5	2105	13	1.26	.95	.97	2.08	(3.28)

Note: % = Percentage, Mnsq = Mean Square

The categories of the CPC ranged from -1.61 to 2.08 and they measured between -2.94 and 3.28 with the average measure not increasing proportionally with the categories (Table 2). The CPC in Figure 1 highlights that the peaks of categories 2 and 3 were not distinctive enough to illustrate further that the thresholds were not well-defined (see Figure 1).

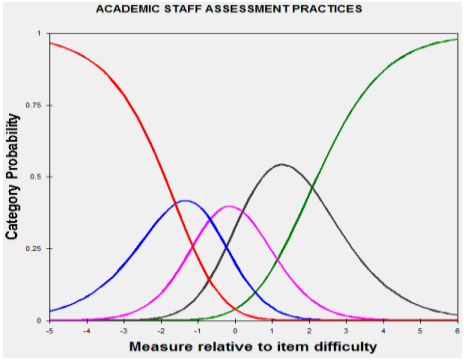

Figure 1. Category Probability Curve for the Un-Collapsed Model

According to the CPC in Figure 1, if the persons' skills in assessing students were less 1 logit than the difficulty of the items, the probability of selecting response category "5" is close to 0, selecting response category "4" is about .10, selecting response category "3" is close to .30, selecting response category "1" is close to .25, and selecting response category "2" is close to .40. The Cronbach's alpha (KR20) of the model was .93, and there were no items with negative point bi-serial correlations. This highlights that all the 50 items in the APIM scale functioned adequately on the latent trait (Linacre, 2006). The infit and outfit statistics analyse the internal and external validity of a measurement instrument respectively (Linacre, 2009). From the item misfit statistics, the mean squares (MnSq) and z-scores (Zstd) values of Item 28 (administering quizzes) were both 1.61 which is greater than the critical value (1.50 logits) to reveal that it is a misfit in the scale, as in Table 3. The MnSq

value of 1 and the z-scores of 0 are ideal for a Likert scale (Bond & Fox, 2007; Linacre, 2009; Wright et al., 1994).

Table 3
Item Misfit Statistics

				Infit		Outfit		PT-Corr
No	Item	Measure	Error	MnSq	Zstd	MnSq	Zstd	
28	Q28	-.03	.07	1.61*	6.7*	1.61*	6.8*	.21

Note: * = Misfit, MnSq = Mean Square, Zstd = Standardised Scores.

Collapsing Categories

Collapsing the categories of the APIM scale was done to improve on the distinctiveness of the categories in the model. During the collapsing of the categories, Item 28 was excluded because inestimable items do not contribute to the construct, have less discriminatory power, and degrade the scale in Rasch measurement analysis (Pomeranz et al., 2008). Category 2 was collapsed into category 1 to have the extreme categories of the scale reserved (not at all skilled and highly skilled). Collapsing the categories led to realising categories which are more distinctive, well-defined, and monotonic in their step ordering calibrations. The CPC of the collapsed categories in Figure 4 show that each category represented a distinct portion of the underlying construct.

In the collapsed model the reliabilities for both items and persons were the same (.93), and did not differ from those of the un-collapsed model (.93). The Cronbach's alpha (KR20) for the 49 items in the APIM scale was .93 (see in Table 4). This was greater than the critical value (.70) indicating good reliability for a measurement instrument (see Bond & Fox, 2007; Fisher, 2007; Garson, 1998; Gleim & Gleim, 2003; Tavakol & Dennick, 2011). The Cronbach's alpha value tentatively explained that APIM scale satisfied evidence of the construct being reliable (Gleim & Gleim, 2003). Despite the categories which were not well-defined in the un-collapsed model, the reliabilities for both the items and persons reveal that the APIM scale was adequate in measuring the skills in assessment practices among university lecturers. The observed reliabilities reflect that there was consistency in the items and the instrument which measured the university lecturers' assessment skills. After collapsing the categories all the 49 items in the model had adequate fit statistics with the peaks of their categories distinctive enough to fit the theoretical curve (see Bond & Fox, 2007).

The item separation index of 3.70 in Table 5 reflects that the items were spread into approximately four levels. This shows that the items were sufficiently separated in terms of their difficulty, and highlighted the direction and meaning of the latent scale (Wright & Stone, 2004). Good item separation is an indication of a small error within the items and evidence of uni-dimensionality in a measurement construct (Bond & Fox, 2001). The persons'

separation index of 3.61 in Table 4 means that the lecturers' assessment skills were split into approximately four levels on the Rasch scale.

Table 4
Summary Statistics of the 303 Measured Lecturers

	Raw Score	Count	Measure	Model Error	Infit MnSq	Infit Zstd	Outfit MnSq	Outfit Zstd
Mean	122.5	49.0	-.07	.18	1.00	-.1	1.00	-.1
S.D	20.	.0	.54	.02	.29	1.7	.30	1.7
Max	182.0	49.0	2.61	.30	1.47	3.1	1.50	3.1
Min	73.0	49.0	-1.80	.17	.51	-2.8	.48	-2.9
Real RMSE	.20	Adj.SD .71	Separation 3.61		Lecturer Reliability .93			
Model RMSE	.19	Adj.SD .71	Separation 3.84		Lecturer Reliability .94			
S.E. of Item Mean = .04								

Note: Cronbach Alpha (KR-20) Lecturers Raw Score Reliability = .93, Deleted: 18 Persons

Table 5
Summary Statistics of the 49 Measured Items

	Raw Score	Count	Measure	Model Error	Infit MnSq	Infit Zstd	Outfit MnSq	Outfit Zstd
Mean	757.4	303.0	.00	.07	1.00	-.1	1.00	.0
S.D	53.8	.0	.29	.00	.17	2.2	.16	2.2
Max	884.0	303.0	.64	.08	1.48	5.3	1.49	5.9
Min	639.0	303.0	-.71	.07	.74	-3.9	.74	-3.9
Real RMSE	.08	Adj.SD .28	Separation 3.70		Item Reliability .93			
Model RMSE	.07	Adj.SD .28	Separation 3.84		Item Reliability .94			
S.E. of Item Mean = .04								

Note: Deleted: 1 Item

The lecturers' assessment skills mean estimate of -.07 in Table 4 indicates that the sample's assessment skills were slightly lower than the mean of items (.00). In good results the person and item averages should correspond to 0 or, be close to 0. The standard deviation in the persons' assessment skills (.54) in Table 4 indicates that there was an average spread of the persons' measures. The average model error is small which is between .07 and .08 to highlight that the items did not greatly deviate from each other in terms of item response. The item mean of 0 indicates an average spread of the items to the participants in the study, while the standard deviation of .29 reflects no greater dispersion of the item measures. This reflects that the items in the APIM scale had an average spread in the item measures on which they purportedly measured, and were appropriate for the sample (see Pomeranz et al., 2008), as seen in Figure 3.

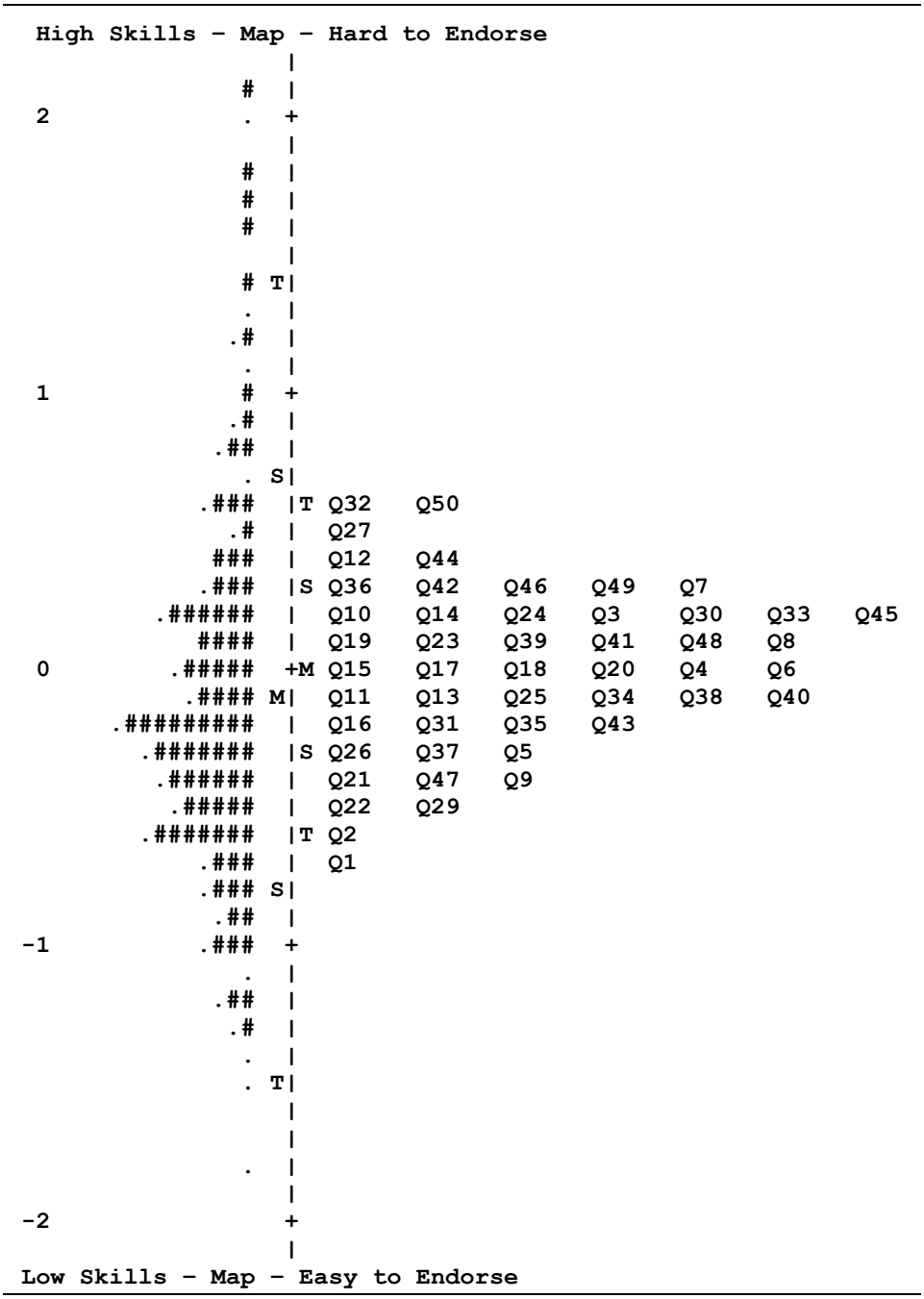

Figure 2. Person – Item Map for 49 Items

According to the person-item map in Figure 2, all the items had skills they measured, but some skills lacked items to measure them. From the eigen-values of the principal component analysis of residuals, it is noted that the item inter-correlations that were not accounted for by the latent trait were not too high for all the contrasts (Table 6). The results in the principal component matrix reveal that the APIM scale is a uni-dimensional instrument because there is good agreement between the empirical and modelled percentages. The ratio of variance explained by the items in the model and the first contrast is 3.8 (32.9/8.7), which is below the critical value (4.0). The variance explained by the measure (52.6%) is also strong to conclude that the uni-dimensionality evidence of the APIM is met.

Table 6
Principle Component Analysis of Standardized Residual Correlation

	Empirical		Modelled
	Eigen values	% of variance	% of variance
Total variance in observations	28.9	100.0	100.0
Variance explained by measures	15.2	52.6	52.7
Variance explained by persons	5.7	19.7	20.0
Variance explained by items	9.5	32.9	33.2
Unexplained variance (total)	13.7	47.4	47.3
Unexplained variance explained by 1st factor	2.5	8.7	
Unexplained variance explained by 2nd factor	2.0	7.0	
Unexplained variance explained by 3rd factor	1.7	5.9	
Variance: By item/ 1st contrast		32.9/8.7 = 3.8	

Note: Last Row: Ratio of Variance Explained by Items to Unexplained Variance in the 1st Contrast

According to the Item Characteristic Curve (ICC) in Figure 3 it is noted that most of the observed scores in the ten-class interval perfectly aligned on the expected curve, and were very close to one another with minimal deviations. This shows that there was good fit between the expected values and observed scores. The category probability curve in Figure 4 highlights that the categories are monotonically ordered, and are more distinctively defined than in the categories before collapsing. This is because the Rasch-Andrich thresholds and the categories are distinctively ordered while predicting the item scores as a function of both the item and person locations on the latent trait (see Hagquist, 2001, 2008).

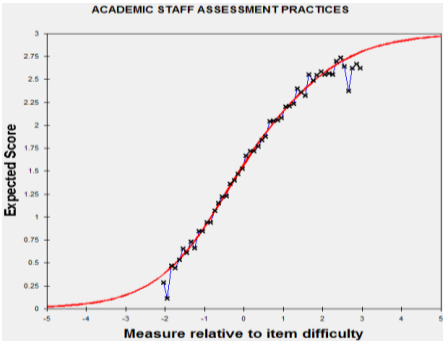

Figure 3. Item characteristic Curve for the 49 items

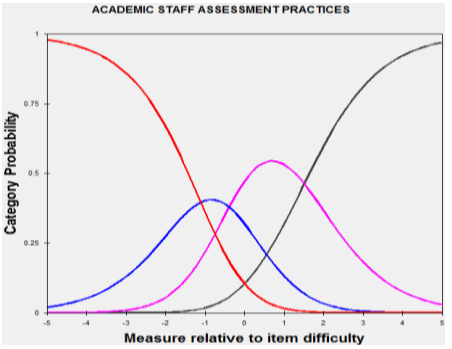

Figure 4. Category Probability Curve for collapsed model

In comparison of the ICC and CPC in Figure 3 and Figure 4 respectively, the university lecturers that had low skills in assessment practices had a higher probability of rating high the items which are easy to endorse in the APIM scale. On the other hand, university lecturers with high skills in assessment practices also had a high probability of endorsing items which were hard to endorse.

Differential Item Functioning

The items of the APIM scale were examined on how they functioned on the different groups of participants who responded to them. According to Bond and Fox (2007) Rasch measurement analysis instruments need to have items with relative estimates, that is, those which are invariant across sub-groups of persons they measure. Examining gender on the differential item functioning, it was discovered that male and female lecturers did not differ on the items in the APIM scale. The differential item functioning difficulty indexes for both male and female lecturers were not different when other elements were kept constant at a confidence level of 95% of the t-value. It is noted that the maximum value of the items for the t-values is 1.84 which is less than the critical level (2.00). The results of the differential contrasts of all the items are observed to be less than .50 which indicates that the items of the APIM scale function appropriately to both male and female lecturers (see Lai & Eton, 2002; Linacre, 2011; Pallant & Tennant, 2007). According to the differential item functioning measures it is revealed that the items do not seem different for the two groups (males and females) of lecturers in terms of functionality as seen in Figure 5.

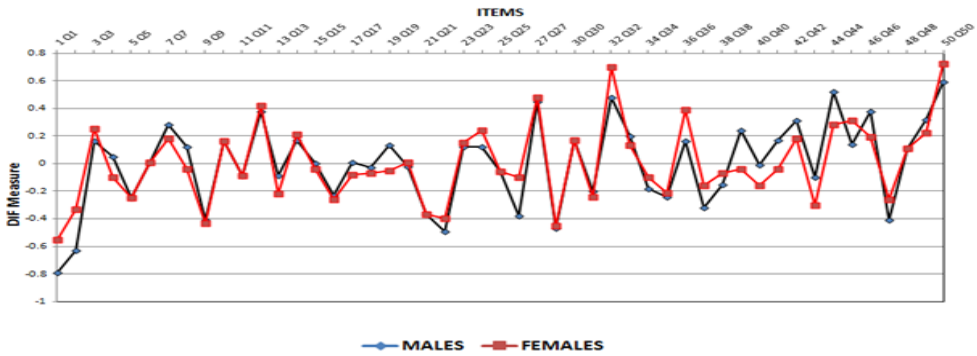

Figure 5. GDIF Plot of Assessment Practices among Lecturers

Discussion and Conclusion

In this study it has been demonstrated that the APIM scale is adequate in measuring assessment practices skills among university lecturers. Only one item was found not function as expected. It was removed from the scale to control quality in item functioning of the remaining items (see Jackson, Draugalis, Slack, & Zachry 2002). Using CTT approach, the Cronbach's alpha (KR20) and the point bi-serial coefficients gave a tentative explanation that the APIM scale is appropriate in measuring assessment practices skills among university lecturers. The IRT using Rasch measurement analysis analysed the fit statistics, reliabilities for persons and items, category functioning, and DIF to conclude that the APIM scale is adequate in measuring assessment practices skills among university lecturers. According to the gender differential item functioning (GDIF) results, the APIM scale is invariant to gender highlighting precision in the measurement. The results of the GDIF in this study do not show measure of the same capabilities (assessment skills) of the different groups of university lecturers (males and females) but only their response differences (see Maller, 2001). From the results of this study the items in APIM scale have been found to be adequate in measuring university lecturers' assessment practices skills. The results of principle component analysis of residuals have demonstrated that the APIM scale is sound and internally consistent (see Andrich, 1988). Though the APIM scale has been found to be a sound instrument in measuring assessment practices skills among university lecturers, it is recommended that there is room for improvement of the scale by adding more items to measure skills which were not captured in this study (see Figure 2). It is also recommended that more validations should be done on the scale to elaborate it further. This would expand its applicability beyond the university lecturers to instructors in other higher education institutions like colleges and polytechnics.

References

Ainol-Madziah, Z., & Noor Lide, A. K. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research, 2*, 1-20.

Ainsworth, L. (2003a). *Power standards: Identifying the standards that matter most.* Englewood, CO: Advanced Learning.

Ainsworth, L. (2003b). *Un-wrapping the standards: A simple process to make standards manageable.* Englewood, CO: Advanced Learning.

Ainsworth, L., & Viegut, D. (2006). *Common formative assessments: How to connect standards based instruction and assessment*. California: Corwin.

Airasian, P. W., & Russell, M. K. (2008). *Classroom assessment: Concepts and applications* (6th ed.). New York: McGraw Hill.

Alkharusi, H. (2012). Educational assessment attitudes, competences, knowledge, and practices: An exploratory study of Muscat teachers in the Sultanate of Oman. *Journal of Education and Learning, 1*(2), 217-232.

Andrich, D. (1988). *Rasch models for measurements.* Newbury Park, CA: Sage Publications.

Andrich, D. (2005). Rasch models for ordered response categories. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopaedia of statistics in behavioural science* (pp. 1698-1707). New York: John Wiley & Sons.

Association of College and Research Libraries [ACRL] (2017). Academic Library impact on student learning and success: Findings from Assessment in Action Team Projects. Prepared by Karen Brown with contributions by Kara J. Malenfant. Chicago: Association of College and Research Libraries.

Biggs, J. (2003). Aligning teaching and assessment to course objectives. *Paper presented at ICHEd Conference: Teaching and Learning in Higher Education: New Trends and Innovations, April 2003*. University of Aveiro, Portugal. Retrieved from http://www.event.ua.pt/iched/main/invcom/p182.pdf

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. New Jersey: Lawrence Erlbaum.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Braney, B. T. (2010). *An examination of fourth grade teachers' assessment literacy and its relationship to students' reading achievement*. (PhD Dissertation). University of Kansas, Kansas.

Brown, G. T. L. (2003), December). Teachers' instructional conceptions: Assessments relationship to learning, teaching, curriculum, and teacher efficacy. *Paper presented at the joint conference of the Australian and*

New Zealand Associations for Research in Education (AARE/NZARE). University of Auckland, Auckland.

Burry-Stock, J. A., & Frazier, C. H. (2008). Revision of the Assessment Practice Inventory Revised (API_R): A combined exploratory factor analysis and polytomous IRT approach. *Paper presented at the American Educational Research Association, March 2008*. New York.

Coates H. (2015) Assessment of learning outcomes. In: A. Curaj, L. Matei, R. Pricopie, J. Salmi, & P. Scott (Eds). *The European higher education area*. Springer, Cham. https://doi.org/10.1007/978-3-319-20877-0_26

Connoley, R. (2004). *Criterion referenced assessment.* Working paper. Geelong-Victoria: Deakin University.

Fatmawati, A. (2011). *Using portfolio to assess student learning of problem solving.* Working paper, Geelong-Victoria: Deakin University.

Fisher, W. P. J. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*, 1095.

Garson, G. D. (1998). *Cronbach's reliability analysis.* North Carolina: North Carolina State University.

Gibbs, G. (2006). Why assessment is changing. In C. Bryan & K. Clegg (Ed.), *Innovative assessment in higher education* (pp. 11-22). London: Routledge.

Gleim, J. A., & Gleim, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert type scales. *In Midwest research practice conference in adult, continuing, and community education*. Ohio: Ohio University.

Greatorex, J., Johnson, M., & Coleman, V. (2017). A review of instruments for assessing complex vocational competence. *Research Matters: A Cambridge Assessment Publication.* Retrieved from: http://www.cambridgeassessment.org.uk/research-matters/

Hagquist, C. (2001). Evaluating composite health measures using Rasch modelling: An illustrative example. *Social and Preventive Medicine, 46,* 369-378.

Hagquist, C. (2008). Psychometric properties of the psycho-somatic problems scale: A Rasch analysis on adolescent data. *Social Indicators Research, 86,* 511-523.

Hagquist, C., & Andrich, D. (2004a). Measuring subjective health among adolescents in Sweden: A Rasch analysis of the HBSC instrument. *Social Indicators Research, 68,* 201-220.

Hagquist, C., & Andrich, D. (2004b). Is the sense of coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling. *Personality and Individual Differences, 36,* 955-968.

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

Hassel, S., & Ridout, N. (2018). An investigation of first-year students' and lecturers' expectations of university education. *Frontiers in Psychology, 8*, 2218. doi:10.3389/fpsyg.

Hindi, N., & Miller, D. (2000). A survey of assessment practices in accounting departments of colleges and universities. *Journal of Education for Business, 75*(5), 286.

Jackson, T. R., Draugalis, J. R., Slack, M. K., & Zachry, W. M. (2002). Validation of authentic performance assessment: A process suited for Rasch modelling. *American Journal of Pharmaceutical Education, 66*, 233-243.

Lai, J. S., & Eton, D. T. (2002). Clinically meaningful gaps. *Rasch Measurement Transactions, 15*(4), 850.

Lai, P. S., Sim, S. M., Chua, S. S., Tan, C. H., Ng, C. J., Achike, F. I., & Teng, C. L. (2015). Development and validation of an instrument to assess the prescribing readiness of medical students in Malaysia. *BMC medical education, 15*, 153. doi:10.1186/s12909-015-0433-z

Linacre, J. M. (1999a). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (1999b). Category disordering vs. Step (threshold) disordering. *Rasch Measurement Transactions, 13*(1), 675.

Linacre, J. M. (2006). *WINSTEPS: Rasch measurement computer program* (version 3.60). Chicago: Winsteps.com.

Linacre, J. M. (2009). *WINSTEPS computer software* (version 3.68). Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2011). *User's guide to WINSTEPS-MINISTEPS*. Rasch model computer programs manual 3.73.0.

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardisation sample. *Education and Psychological Measurements, 61*(5), 793-817.

Martell, K., & Calderon, T. (2005). Assessment of student learning in business schools: What it is, where we are, and where we need to go next. In K. Martell & T. Calderon (Eds.), *Assessment of student learning in business schools: Best practices each step of the way* (pp. 1-22). Tallahassee, Florida: Association for Institutional Research.

Marzano, R. J. (2000). *Transforming classroom grading.* Alexandria, VA: Association for Supervision and Curriculum Development.

McClarty, K. L., & Gaertner, M. N. (2015). *Measuring mastery: Best practices for assessment in competency-based education*. Center for College & Career Success, Pearson. Retrieved from: https://www.luminafoundation.org/files/resources/measuring-astery.pdf

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32.

Messick, S. (1989). Validity. R. In Linn (Ed.), *Educational measuremen*t (3rd ed.). Washington, D.C: American Council on Education.

Ministry of Education [MOE]. (2010). *Assessment and evaluation in Saskatchewan*. Saskatchewan: MOE, Assessment and Evaluation.

National Council for Curriculum and Assessment [NCCA]. (2005). *Supporting assessment in schools-1: Assessment in Primary Schools*. Retrieved from http://www.ncca.ie/uploadedfiles/primary/ass_pr_schs_05.pdf

O'Connor, K. (2009). Reforming grading practices in secondary schools. *Principal's Research Review, 4*(1), 1-7.

O'Donovan, B., Price, M., & Rust, C. (2001). The student experience of criterion referenced assessment through the use of a common criteria assessment grid. *Innovations in Learning and Teaching International, 38*(1), 74-85.

O'Grady, G. (2006). *Guide for good practice in assessment*. Ngee Ann Polytechnic: Singapore.

Orzolek, D. C. (2006). The paradox of assessment: Assessment as paradox. *Research and Issues in Music Education, 4*(1), 1-5.

Pallant, J. F., & Tennant, A. (2007). An Introduction to Rasch measurement model: An example using the hospital anxiety and depression scale (HADS). *Journal of Clinical Psychology, 46*, 1-18.

Peterson, M. W., Einarson, M. K., Augustine, C. H., & Vaughan, D. S. (1999a). *Institutional support for student assessment: Methodology and results of a national survey*. Stanford, CA: Stanford University, NCPI.

Peterson, M. W., Einarson, M. K., Augustine, C. H., & Vaughan, D. S. (1999b). *Designing student assessment to strengthen institutional performance in comprehensive institutions*. National Canter for Postsecondary Improvement. Stanford, CA: Stanford University, NCPI.

Pomeranz, J. L., Byers, K. L., Moorhouse, M. D., Velozo, C.A., & Spitznagel, R. J. (2008). Rasch analysis as a technique to examine the psychometric properties of a career ability placement survey subtest. *Rehabilitation and Counselling Bulletin, 51*(4), 251-259.

Popper, E. (2005). Learning goals: The foundation of curriculum development and assessment. In K. Martell, & V. Calderon (Eds.), *Assessment of student learning in business schools: Best practices each step of the way* (pp. 1-23). Tallahassee, Florida: Association for Institutional Research.

Price, M., & Rust, C. (1999). The experience of introducing a common criteria assessment grid across an academic department. *Quality in Higher Education, 5*(2), 133-44.

Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests.* Copenhagen: Institute of Educational Research.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*(1), 49-57.

Reeves, D. B. (2004). The Case against the Zero. *Phi Delta Kappan, 86,* 324-326.

Rousselot, N., Tombrey, T., Zongo, D., Mouillet, E., Joseph, J. P., Gay, B., & Salmi, L. R. (2018). Development and pilot testing of a tool to assess evidence-based practice skills among French general practitioners. *BMC Medical Education, 18,* 254 https://doi.org/10.1186/s12909-018-1368-y

Russell, J., & Markle, R. (2017). Continuing a culture of evidence: Assessment for improvement. *Educational Testing Service Research Report Series, 1*, 1-10. https://doi.org/10.1002/ets2.12136

Satterly, D. (1989). *Assessment in schools* (2nd ed.). Oxford: Blackwell.

Shaw, F., Wright, B., & Linacre, J. M. (1992). Disordered steps? *Rasch Measurement Transactions, 6*(2), 225.

Singh, C. K. S., Lebar, O., Kepol, N., Abdul-Rahman, R., & Mukhtar, K. A. M. (2017). An Observation of classroom assessment practices among lecturers in selected Malaysian higher learning institutions. *Malaysian Journal of Learning and Instruction, 14*(1) 23-61.

Smith, E. V. Jr., Wakeky, M. B., de Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy and other. *Research, Educational and Psychological Measurement, 63*(3), 369-391.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2,* 53-55.

Watkins, D. (1998). Assessing university students in Hong Kong: How and why? In D. Watkins, C. Tang, J. Biggs, & R. Kuisma (Eds.), *Assessment of university students in Hong Kong: How and why, assessment portfolio, students' grading* (pp. 5-27). Hong Kong: University of Hong Kong.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement, 1,* 83-106.

Wright, B. D., & Stone, M. H. (2004). *Making measures.* Chicago: Phaneron Press.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin, L. P. (1994). Reasonable mean square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Zhang, Z., & Burry-Stock, J. A. (1994). *Assessment practices inventory.* Tuscaloosa, AL: The University of Alabama.

Appendix: Assessment Practices Inventory Modified Scale

S/N	Items	NS	LS	SW	S	HS
1	Communicating grading procedures to students/ parents.	1	2	3	4	5
2	Marking according to university assessment policy and standards.	1	2	3	4	5
3	Considering rubrics when marking/ grading students in an assessment.	1	2	3	4	5
4	Using assessment results to enhance instructional delivery.	1	2	3	4	5
5	Protecting students' confidentiality with regard to test scores or grades.	1	2	3	4	5
6	Aligning tests to university assessment standards.	1	2	3	4	5
7	Developing standardised proficiency tests.	1	2	3	4	5
8	Demonstrating 'good' or 'bad' projects when I give assessments.	1	2	3	4	5
9	Providing students with guidelines to assess their own work.	1	2	3	4	5
10	Using previous assessment results to improve course outlines.	1	2	3	4	5
11	Using assessment results in making decisions about individual students (graduation, research grants, placement etc).	1	2	3	4	5
12	Using assessment results to prepare content for the next lecture(s).	1	2	3	4	5
13	Awarding scores according to the weight of the project, test, or item being assessed.	1	2	3	4	5
14	Communicating assessment results to students.	1	2	3	4	5
15	Using the criterion-referenced assessment/ grading model.	1	2	3	4	5
16	Incorporating extra credit activities in student grades.	1	2	3	4	5
17	Involving students in assessment design and development.	1	2	3	4	5
18	Recognizing unethical use of assessment information.	1	2	3	4	5
19	Recognizing inappropriate methods of assessment.	1	2	3	4	5
20	Selecting an appropriate assessment that can accurately measure students' achievement.	1	2	3	4	5
21	Organizing student assessment results to make meaning.	1	2	3	4	5
22	Conveying to students what subject matter is to be assessed.	1	2	3	4	5
23	Using cognitive taxonomies (e.g. Quellmalz, Bloom etc) in designing assessment.	1	2	3	4	5
24	Constructing tests items that match classroom instruction.	1	2	3	4	5
25	Avoiding teaching to test when preparing students for assessments.	1	2	3	4	5
26	Calculating and interpreting student assessment scores.	1	2	3	4	5
27	Incorporating effort into students' final achievement grades.	1	2	3	4	5
28	Administering quizzes.	1	2	3	4	5
29	Constructing tests based on clearly defined course objectives.	1	2	3	4	5

(continued)

Appendix: Assessment Practices Inventory Modified Scale (continued)

S/N	Items	NS	LS	SW	S	HS
30	Comparing obtained assessment results with instructional and course objectives.	1	2	3	4	5
31	Emphasizing quality control when assessing students.	1	2	3	4	5
32	Writing multiple-choice questions.	1	2	3	4	5
33	Writing true/false questions.	1	2	3	4	5
34	Writing fill-in-the-blank/short answer questions.	1	2	3	4	5
35	Writing essay questions.	1	2	3	4	5
36	Writing test items at higher cognitive levels (inference and evaluation).	1	2	3	4	5
37	Constructing marking schemes for scoring essay questions.	1	2	3	4	5
38	Ensuring adequate content sampling when setting tests.	1	2	3	4	5
39	Using previous assessment results in designing the next test(s).	1	2	3	4	5
40	Using assessment results to define/ refine a grading rating scale.	1	2	3	4	5
41	Communicating assessment criteria to students in advance.	1	2	3	4	5
42	Controlling external stakeholders influence when assessing students.	1	2	3	4	5
43	Assessing group/ individual class participation.	1	2	3	4	5
44	Controlling institutional (university) influence when assessing students.	1	2	3	4	5
45	Incorporating students' attendance in academic achievement (Grades).	1	2	3	4	5
46	Weighing and interpreting grades from assessments.	1	2	3	4	5
47	Assigning grades.	1	2	3	4	5
48	Using overall results of students' performance (e.g. Grades, CGPA etc).	1	2	3	4	5
49	Revising tests based on item analysis.	1	2	3	4	5
50	Obtaining diagnostic information from assessment results.	1	2	3	4	5

1 = not at all skilled (NS), 2 = a little skilled (LS), 3 = some-what skilled (SW),
4 = skilled (S), 5 = highly skilled (HS)