Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Developing and validating instrument of alignment understanding with school assessment practice

Eftah Moh Abdullah *, Abd Aziz Abd Shukor, Norazilawati Abdullah, Mohammad Aziz Shah Mohamed Arip

Faculty of Education and Human Development, Sultan Idris Education University, Tanjung Malim, Malaysia

ARTICLE INFO

Article history:
Received 17 August 2016
Received in revised form
20 October 2016
Accepted 30 October 2016

Keywords:
Assessment understanding
Assessment practice
Instrument development
Validation
Alignment

ABSTRACT

This study aims to develop and validate the Understanding Alignment with School Assessment Practice instrument (2KAPS). The instrument consists of 27 items. The 2KAPS questionnaire validation involves 109 teachers who taught Form 1 and Form 2 students (teachers directly involved in the School Assessment Practice implementation) in one district in Perak. The instrument was developed in several stages such as building the understanding alignment model and an assessment practice generated based on alignment models from the literature review, determining the main constructs in the assessment expectations, determining the chosen practice in line with the assessment expectations, the use of the Likert scale with three categories (Full agreement=3, Lack of agreement=2 and No Agreement=1) which indicated that there was an alignment between the assessment practice and the teachers' assessment understanding, acquiring the content validity from experts and the analysis of items using the Rasch Measurement Model. The instrument validity and reliability had been conducted by identifying the Rasch fit statistics, item difficulty, unidimensionality, item reliability as well as 2KAPS item map. The Rasch analysis showed that the item reliability was valued at 0.92 while the Cronbach Alpha value was 0.90. All the items fit the model as their MNSQ values were between 0.7 and 1.35. The dispersion of items from 2KAPS data was 3.29 which indicated the existence of 3 to 4 item strata. No item showed a negative point measure correlation or less than 0.2 and this generally indicated that the item discrimination was very good. The data showed that the mean for a person was measured at 1.19 logits with a standard deviation of 1.12 logits while the item mean value was zero with a standard deviation of 0.52. This indicates that the position of item and person does not fully match and thus shows a medium difficulty. The overall item quality was good and all 27 items of 2KAPS were retained.

1. Introduction

The implementation of School-based Assessment in Malaysia had given rise to a lot of issues which may show different understanding and readiness from various parties involved. The School-based Assessment had been implemented in Queensland, Australia but the assessment accountability issues still remained a major problem in the implementation (Klenowski and Valentina, 2011). The teaching and learning process would be affected if the assessment was not in line with the curriculum (Boss et al., 2001) and the weak alignment between what is taught with what is assessed will in turn affect students' achievement. The School-based assessment as described by the Ministry of Education, Malaysia, is a classroom assessment which would enable the transformation of the assessment practice from a post-teaching assessment to a pre-, while and post-teaching assessment, as well as moving away from judging students to guiding them and the production of more information-based learning evidence. The function of the teacher as someone who merely teaches would shift to enable students to utilize the holistic achievement and potential to achieve success. The concept of assessment as delivered in Malaysia's education assessment transformation shows a paradigm shift from an assessment based on competition and judgment to a vision of assessment

* Corresponding Author.
Email Address: eftah.a@fppm.upsi.edu.my (F. M. Abdullah)
https://doi.org/10.21833/ijaas.2016.11.004
2313-626X/

as teacher reflection which would improve teaching and learning. School-based assessment does not use the students as a means of comparison; instead, it aims to evaluate students fairly based on their abilities, skills, talents and potential. As such, the school-based assessment implemented means a paradigm shift in line with the assessment concepts stated by Huba and Freed (2000) which include collecting and discussing information from various sources to build in-depth understanding of what the students know, understand and are able to carry out, based on their learning experience. The importance of alignment becomes easier to grasp when teachers understand the alignment between the learning outcomes, strategies and teaching and learning activities, as well as the suitable assessment for a particular assignment in a learning activity consistent with the desired learning objectives. An aligned system will utilize limited resources effectively. The alignment of learning objectives and assessment of learning outcomes will enable those involved in the field of education to work towards the same goal. Spady (1994) defined 'alignment' as 'a matching exercise'. The alignment between understanding and the assessment practice is a perfect match of important aspects involving the assessment understanding expectations with the assessment practice. Alignment is the degree whereby the assessment understanding expectations and assessment practice correspond and contribute towards the continuity of one another as a guide of what is expected to happen. The Understanding Alignment with School Assessment Practice **(**2KAPS) instrument was developed to identify whether the understanding expectations and the assessment practice in the classroom were aligned. The development and validation of the 2KAPS instrument was conducted in several important stages. The aim of this study was to utilize the Rasch Model to identify the instrument validity and reliability. The Rasch analysis was conducted in 6 steps:

1. Rasch Fit Statistics
2. Item difficulty measurement
3. Item polarity
4. Unidimensionality
5. Dispersion and Reliability
6. 2KAPS item map

2. Literature review

Validity refers to the suitability of inference acquired from the information of the assessment outcome (Cronbach and Thorndike, 1971). Sometimes, alignment may also be related to the validity of a test. Alignment also refers to the extent of which an element of system works together to guide teaching and learning, with the students' learning as the ultimate aim (Moss, 1992). Two or more systems are aligned if they correspond to one another. Expectations can be understood as what the teacher should know about assessment and what they can do with the knowledge of the assessment. Expectations can be defined in various ways, especially in terms of the expectations of the learning outcomes to be measured via assignment or tests. The main aim of classroom assessment is to collect information about the students' teaching (McMillan, 2007). Assessment does not only include pencil and paper tests, but it also involves the retrieval of information about the students, which involves questionnaires, interviews, presentations of portfolios, etc. The learning environment may also present a suitable space for students to be assessed or a place where they can ask about their learning objectives. According to Biggs (1996), there are four main steps which could be utilized as guidelines in building a constructive alignment:

1. Defining the learning outcomes
2. Choosing teaching and learning activities which would direct towards the fulfillment of learning outcomes
3. Assessing the learning outcome to identify whether it is aligned with targeted outcome
4. Developing meaningful information based on assessment information whether in qualitative or quantitative form which would be more useful in supporting the students' achievement.

To summarize the elements suggested by Biggs (1999), learning objectives can be achieved when there is alignment between curriculum, teaching, learning and also assessment. If there is no emphasis on alignment, it would be difficult to guide students and to implement the curriculum as planned.

The School Assessment Expectations Understanding refers to the main features of the education assessment system transformation, which include a holistic assessment system, flexible, standards-based and forms a part of the teaching and learning process. The Assessment Practice, on the other hand, comprises the teachers' assessment practice in the classroom. The main features in the education system must work together to deliver a process which would be targeted in the same direction, as well as create the drive towards an effective assessment system transformation. The educators admit that if the elements of the policy are not aligned, the system would be fragmented, causing a mix-up in the information process, which in turn makes the system less effective (Newman, 1993).

Assessment understanding could be defined as what the teacher should know about assessment and what they can do using the assessment information. Assessment refers to the procedure in the system utilized by the teacher to grade, identify students' needs, provide motivation, identify weaknesses in teaching and improve teaching to become more effective (Ohlsen, 2007). Webb (1997) stated that the shared attributes between expectations and assessment are:

1. The categories of the same content should be in both the expectations and assessment.

2. Expectations and assessment should require students to know the information at the same level, able to move/utilise knowledge in different contexts and have the same basic knowledge.
3. Expectations and assessment should comprise topics and ideas in slightly similar categories.
4. Expectations and assessment should match and be similar in terms of basic concepts and knowledge of the concept's meaning.
5. Expectations and assessment should stress on the topic's content, activity and teaching tasks.
6. Expectations comprise more than concepts, procedures and applications, such as helping to develop attitudes, beliefs, a wider vision etc.

Webb (1997) also outlined assessment criteria congruent with expectations such as below:
i. An assessment to evaluate students utilizes various forms of assessment across several domains such as knowledge, character and performance.
ii. Rubrics or criteria to define whether the teacher has succeeded in assessing the achievement and used for evaluating students' work.
iii. A fair evaluation based on continuous assessment
iv. The quality of assessment system can be used to reinforce teaching.

The implementation of classroom assessment is not an easy task as it comprises many activities such as building pencil and paper tests, measuring achievement, grading, interpreting test scores, communicating the assessment results and using the assessment results to make inferences about teaching and learning. Stiggin (2002) lists seven important competencies to be mastered by the teacher:

1. Linking assessment with its aim clearly
2. Defining clearly the students' achievement expectations
3. Using suitable assessment methods
4. Avoiding bias in the assessment
5. Communicating effectively about the students' performance
6. Using assessment as intervention in teaching and learning

Although the standards have been determined, teacher competency is the basic competency which could be used to direct teachers towards a more effective assessment system.

Alignment refers to the extent of which expectations and assessment match and contribute to each other's survival as a guide towards what is expected to happen (Webb, 1997). Assessment understanding can be understood as what the teacher should know about assessment and what he or she could achieve with that knowledge. It refers to the procedures in the system utilized by the teacher to grade, identify students' needs, provide motivation, identify weaknesses in teaching and improve teaching to become more effective (Ohlsen, 2007). The main aim in classroom assessment is to gather information about students' teaching (McMillan, 2007).

3. Methodology

The instrument was developed in various stages as stated below:

1. Determining the main constructs in assessment understanding expectations.
2. Generating the model of the alignment of expectations and the assessment practice based on the models of the alignment between assessment practice and expectations understanding, which include items built from the five main criteria suggested in the study by Webb (1997) regarding the criteria for alignment of expectations and assessment in the teaching of science and mathematics.
3. Developing the items after conducting the literature review.
4. Checking the items in a workshop conducted with the researchers and the questionnaire items would be evaluated by 10 teachers from 2 schools in the Batang Padang district. The teachers would comment on the understanding element in each item and this is done to provide face validity. The instrument consists of five main constructs focusing on the content of assessment practice. The instrument consists of five main constructs focusing on assessment practice content which contains sub-constructs of understanding aspect relating to school assessment, consistency of the knowledge depth about the assessment, the range of knowledge used to describe the performance of students, the comparison of knowledge structure, balanced representation and consonant difference. The second construct is the circulation across age and grade with the sub-construct of the best cognitive determined through studies and understanding. The third sub-construct involves transparency and fairness with the sub-construct of information transparency. The fourth construct is the pedagogical implication with sub-constructs which are students' involvement and effective classroom practices, effective evaluation and usage of technology, resource and materials. The fifth construct is the system usability.
5. Content validity is examined by two experts in the field of testing and evaluation.

The Alignment of Understanding Expectations with School Assessment Practice questionnaire (2KAPS) utilizes the Likert Scale with three categories (Full agreement=3, Lack of agreement=2 and No Agreement=1) to show the agreement or the alignment between understanding expectation and teacher assessment practice. The 2KAPS questionnaires were distributed to 109 teachers in the one district in Perak, Malaysia. The sample size of 109 teachers teaching in Form 1 and Form 2 and 27 2KAPS items could be considered as able to produce a stable index. This is because the Rasch

analysis requires a sample of 100 respondents and 20 items for the data to be considered stable (Green and Frantom, 2002).

4. Result

4.1. Rasch fit statistics

Bond and Fox (2003) described that the item fit estimate would provide information on the pattern of distribution of item difficulty and whether it approaches a certain model or otherwise. Linacre (1994) suggested that the mean squared value (MNSQ) fit for the model would be INFIT and OUTFIT which are between the scale of 0.6-1.4; the value of 1.4 shows a variability of more than 40% while a value of 0.6 shows a variability of less than 40% as expected by the Rasch model. The Rasch model via the INFIT and OUTFIT statistics is exemplified in two ways which are MNSQ and standard mean squared (Zstd). MNSQ shows how far the data's random response pattern fits in with the model. This shows the difference in magnitude between the expected response and the monitored response. Bond and Fox (2003) explained that the item fit estimate would give information on the distribution pattern of item difficulty and whether it approaches a certain model or otherwise. The fit estimate in the mean square value would be used as a method of control so that the data acquired would be consistent with the model.

Table 1 shows a mean squared value which is less than the value of 1, which indicates a lack of variation from the model. The INFIT squared mean (MNSQ) is the ratio between the observed sample variance and the expected variance from the model. This provides evidence of how far the data fits the Rasch model; MNSQ which is less than 1 shows that the student response was closer to the *'Guttman style response string'* (true for all easy items and false for difficult items). Table 1 show that the MNSQ item laid between 0.7 and 1.35. Bond and Fox (2003) stated that the data fits the model based on the MNSQ INFIT value range of 0.6 to 1.4.

4.2. Item difficulty

Item difficulty can be defined using the variable continuum from easy to more difficult as measured using logit units. The item validity is defined via the assessment of item difficulty; all the items are arranged in a hierarchical position to define each construct. The arrangement of the 2KAPS item difficulty is shown in Table 1. The instrument validity according to the Rasch model is the construct validity/idea or the order of items (Smith and Miao, 1994; Wright and Stone, 1979). Usually the mean of an item is considered as zero in the Rasch model (Bond and Fox, 2001). If the item measure and the ability of an individual match closely, the item would provide a lot of information about the individual and this is known as a latent trait. The entire 2KAPS test can be 'targeted' if the mean of an individual falls in the range of 2 standard deviations from the mean. The target for the data acquired was sufficient as the highest measurement (item 12) was 1.07 (in the range of 2 standard deviations) whiles the lowest measurement in 2 standard deviations. All the items showed a positive was item 4 with a value of -1.21 (still in the range of 2 standard deviations). All the items fit the model as their MNSQ values were between 0.7 and 1.35.

4.3. Item polarity

All the items showed positive item discrimination and a pattern which showed a high validity via a positive correlation point size value. Point Measure Correlation is a statistical item used to show the correlation results between one points (a response choice) with a continuous variable (scores for all candidates in a test). Point Measure Correlation in Rasch statistics uses the mean square value of the residual item which is sensitive to the items which have failed to relate to the test scores and point-biserial items with very large values. This means that the correlation point size in Rasch statistics is sensitive to the interaction of items which do not follow a certain model in the calibration sample (Wright and Stone, 1979). Pray and Popovich (1885) stated that an acceptable critical point measure correlation of an item is 0.2 or more. Masey (1995) stated that a discrimination index of less than 0.2 is weak, while an index more than 0.4 is good. Table 1 show that the 2KAPS item had a lowest point measure correlation of 0.35. No item showed a negative point measure correlation or less than 0.2 and this generally indicated that the item discrimination was very good.

4.4. Unidimensionality

Unidimensionality and local independence in the Rasch analysis are assessed by the FIT item statistics. The unidimensionality and local independence are achieved when a set of FIT statistics criteria with the model has been fulfilled and the item reliability index has also been fulfilled. Unidimensionality and local independence both provide empirical evidence for detecting:

i. When an item measurement shows a different dimension
ii. When the item is not understood
iii. When the response shows the students' guesswork or special skills

Unidimensionality is an important factor to be considered when using the Rasch model. Unidimensionality can also be assessed using the Principal Component Analysis (PCA). Smith (2002) used an independent t-test to compare teacher location estimate based on a different subset of items. If the deviance from the unidimensionality is small, then the number of different teacher locations from two different sets is also small. The item correlation matrix in PCA is based on the residual or

the difference between what is observed with what is expected to identify other dimension potential. Linacre (2006) found that a latent dimension variance trait of 60% or more as observed is good, while Linacre and Fisher (2012) found that the sum of variance explained as 50 to 60 percent is good enough for a test's quality. The first contrast also functions to show whether it has a sufficient-sized variance to indicate that more than one dimension exists.

Table 1: Item measure, MNSQ (INFIT, OUTFIT) and Point Measure Correlation

Item	Measure	Standard error	MN SQ INFIT	MN SQ OUTFIT	ZSTD	Pt MEA COrr
i12	1.07	.14	.97	.99	-.1	.50
i26	.81	.14	.97	.99	-.1	.60
i21	.68	.14	.66	.76	-2.4	.54
i16	.56	.14	1.03	.98	-.1	.57
i14	.42	.14	.87	.87	-1.2	.51
i8	.41	.14	.96	.95	-.4	.51
i19	.30	.14	.93	.94	-.5	.51
i24	.29	.14	.93	.94	-.5	.51
i3	.25	.14	1.13	1.05	.5	.54
i10	.25	.14	.89	.88	-1.1	.53
i6	.21	.14	.89	.87	-1.1	.55
i1	.21	.14	1.33	1,35	2.7	.42
i25	.17	.14	1.25	1.2	1.6	.41
i17	.12	.14	1.35	1.27	2.1	.52
i7	.07	.14	.96	.96	-.3	.49
i15	-.06	.15	1.05	.00	,0	.50
i11	-.15	.15	.79	.86	-1.1	.52
i2	-.24	.15	1.00	1.01	.1	.47
i9	-.24	.15	.70	.74	-2.1	.53
i20	-.25	.15	.79	.79	-1.6	.55
i18	-.28	.15	1.04	1.15	1.1	.45
i13	-.56	.16	.86	.81	-1.3	.49
i22	-.66	.16	1.11	1.69	3.7	.40
i23	-.66	.16	1.05	.97	-.1	.44
i27	-.68	.16	.95	.91	-.5	.48
i5	-.81	.16	1.11	1.48	2.6	.35
i4	-1.21	.18	1.01	.92	-.31	.37
Mean.	.00	.15	1.00	1.02	.1	
SD	.52	.01	.17	.22	1.5	

Linacre (2007) stated that an easy way to check PCA based on the residual is to ensure that the first contrast has a strength of at least 3 items measured using the eigen value and represents more than 5% of the unexplained variance. The assessment of measured dimension strength as based on Linacre (2006) is that for an explained variance, a measurement higher or equal to 40% is considered a strong dimension, higher or equal to 30% is considered a moderately strong dimension while higher or equal to 20% is considered a moderate dimension. The 2KAPS findings in Table 2 inform us that the measured dimension was 29.0% and this is closer to a moderately strong dimension. 9.1% of the variance could be explained by the first residual contrast which was more than 5% of the unexplained variance (Linacre, 2007). The ratio of 29.0 with 9.1 is 3:1 which points towards a unidimensionality feature. The Cronbach alpha value of 0.90 also indicates a very good unidimensionality.

4.5. Dispersion and reliability

Fit statistics also enable the researcher to detect whether each item contributes to the measure of each construct. The item reliability value can provide an indication whether the items or cluster of items interact well with one another to describe the same attributes (Wright and Stone, 1979). A person's reliability is explained on a scale of 0 to 1 and this provides meaning just like the alpha Cronbach value. Dispersed items and people are calibrated. The dispersed item, people and reliability are used to assess the rate of dispersion across the trait continuum. This measures the dispersion of both item and people in standard unit.

It shows the number of dispersed levels for item and people. The instrument dispersion to be utilized should reach the value of 1; a high dispersion level shows that there is item and person dispersion further along the continuum. A small dispersion value indicates that there may be overlapping items and less person variability in the trait. Dispersion is used to describe how a strata of latent traits could be found using item measurement (Full agreement=3, Lack of agreement=2 and No Agreement=1). Expected dispersion should reach the value of 2.0 to describe all 3 strata. Table 3 shows that the dispersion of items from 2KAPS data was 3.29 which indicated the existence of 3 to 4 item strata while

teacher or person dispersion was 2.49 which showed the existence of 3 people strata. Dispersion indicates reliability. Dispersion reliability for people generally is similar with alpha Cronbach's which shows an instrument's internal consistency reliability. Item reliability was valued at .92 and teacher reliability was valued at .90. Linacre (2007) stated that a reliability value of more than 0.8 showed very good reliability.

Table 2: Standard residual variance (In Eigen Value Units)

	Empirical	(%)		Model
Number of raw variance in observation	38.0	100%		100.00%
Raw variance explained by measurement	11.0	29.0%		29.20%
Raw variance explained by the teacher	5.1	13.40%		13.40%
Raw variance explained by the item	6	15.70%		15.70%
Number of unexplained raw variance	27	71.0%	100%	70.8%
Unexplained variance in the first contrast	2.5	6.5%	9.1%	
Second	2.0			

Table 3: Summary of item and person measure

Summary of person measurement						
		INFIT		OUTFIT		
	Measurement	MSQ	ZSTD	MSQ	ZSTD	Model error
Mean	1.19	1.01	-.1	1.02	-.1	
SD	1.12	.37	1.7	.45	1.7	
Dispersion	2.49					
Reliability	.90					
Summary of Item Measurement						
Mean	.00	1.00	-.1	1.02	.1	.15
SD	.52	.17	1.7	.22	1.5	.01
Dispersion	3.29					
Reliability	.92					

4.6. Item distribution map

This map shows the distribution of people/person and item on the same measurement scale. The scale measures constructs vertically with the most capable person and the most difficult item is placed at the top. The column on the left shows the measure of the person's capability in logits. Table 1 also shows that the item distribution map enables researchers to observe the item function and and teachers' overall capability measurement. To assess item distribution, items need to be measured as less than -2 logits to + 2 logits. A standard error of 0.15 logits is sufficient to indicate that items are different. The empty value between items, if it is more than 0.15 logits, shows that the items differ between one another while an empty value more than 0.30 have to be filled with other items. Table 3 shows that the standard error between items was 0.15 which shows that item dispersion existed and the items differed between one another; on the other hand Diagram 1 shows the arrangement of variables in the 2KAPS instrument continuum. After the calibration had been conducted, some items did not show a perfect match between overall personal ability and item. Table 3 showed that the mean for a person was measured at 1.19 logits with a standard deviation of 1.12 logits while the item mean value was zero with a standard deviation of 0.52. This indicates that the position of item and person does not fully match and thus shows a medium difficulty.

A total of 12 items as illustrated in Table 1 showed items located above the mean item and items which are considered difficult to agree in relative. Table 1 showed Item 12 was the most difficult item whereby the teachers appeared to have difficulty in agreeing that the assessment tasks involved the transfer of information to a new situation. A total 17 items as illustrated in Table 1 showed items located below the mean item (arbitrary mean) and items are considered easy to agree in relative.

5. Conclusion

This study aims to establish an instrument designed to assess Alignment Understanding with School Assessment Practice among teachers. The method used to determine the quality of the instrument has gone through several processes of development also validation of empirical data analysis using the Rasch measurement model. The analyses revealed that all the items fit the model as their MNSQ item laid between 0.7 and 1.35. 2KAPS item had a lowest point measure correlation of 0.35. No item showed a negative point measure correlation or less than 0.2 and this generally indicated that the item discrimination was very good. The Cronbach alpha value of 0.9 also indicates a clear unidimensionality. Item reliability was valued at .92 and teacher reliability was valued at .86 that a value of more than 0.8 showed very good reliability. A dispersion of 0.15 logits is sufficient to indicate that items are different. A total of 12 items located above the mean item and items which are considered difficult to agree in relative. Item 12 was

the most difficult item whereby the teachers appeared to have difficulty in agreeing that the assessment tasks involved the transfer of information to a new situation.

References

Biggs J (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3): 347-364.

Biggs J (1999). Teaching for Quality Learning at University. SRHE and Open University Press, Buckingham, UK.

Bond TG and Fox CM (2003). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Journal of Educational Measurement, 40(2), pp.185-187.

Boss T, Endorf D and Duckendahl C (2001). Informing state assessment from the local level: A district's reflections. Annual Meeting of the Mid-Western Education Research Association, Chicago, Illinois, USA.

Cronbach LJ and Thorndike RL (1971). Educational measurement. Test Validation, American Council in Education, Washington, DC, USA: 443-507.

Green KE and Frantom CG (2002). Survey development and validation with the Rasch model. In International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, USA.

Huba ME and Freed JE (2000). Learner centered assessment on college campuses: Shifting the focus from teaching to learning. Community College Journal of Research and Practice, 24(9): 759-766.

Klenowski V (2011). Assessment for learning in the accountability era: Queensland, Australia. Studies in Educational Evaluation, 37(1): 78-83.

Linacre JM (1994). Sample Size and Item Calibration Stability. Rasch Measurement Transactions, 7 (4): 328.

Linacre JM (2006). Data variance explained by Rasch measures. Rasch Measurement Transactions, 20(1): 1045.

Linacre JM (2007). Standard errors and reliabilities: Rasch and raw score. Rasch Measurement Transactions, 20(4): 1086.

Linacre JM and Fisher WP (2012). Harvey Goldstein's objections to Rasch measurement: A response from Linacre and Fisher. Rasch Measurement Transactions, 26(3): 1383-1389.

Massey A (1995). Evaluation and analysis of examination data: Some guidelines for reporting and interpretation. UCLES Internal Report, Cambridge, UK.

McMillan JH (2007). Formative classroom assessment: The key to improving student achievement. Formative Classroom Assessment: Theory Into Practice, Teachers College Press, New York, USA: 1-7.

Moss PA (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62(3): 229-258.

Newmann FM (1993). Beyond Common Sense in Educational Restructuring The Issues of Content and Linkage. Educational Researcher, 22(2): 4-22.

Ohlsen MT (2007). Classroom assessment practices of secondary school members of NCTM. American Secondary Education, 36(1): 4-13

Pray WS and Popovich NG (1985). The development of a standardized competency examination for doctor of pharmacy students. American Journal of Pharmaceutical Education, 49(1): 1-9.

Smith LI (2002). A Tutorial on Principal Components Analysis. Available online at: http://reflect.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf; 2002.

Smith RM and Miao CY (1994). Assessing unidimensionality for Rasch measurement. Objective Measurement: Theory into Practice, 2: 316-327.

Spady WG (1994). Outcome-Based education: Critical issues and answers. American Association of School Administrators, Arlington, USA.

Stiggins RJ (2002). Assessment crisis: The absence of assessment for learning. Phi Delta Kappan, 83(10): 758-765.

Webb NL (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No. 6, National Institute for Science Education Publications, Washington, USA.

Wright BD and Stone MH (1979). Best Test Design (Rasch Measurement Series). MESA Press, Chicago, USA.