International Online Journal of Educational Sciences, 2012, 4 (1), 91-106

iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Investigating Assessment Practices of In-service Teachers

See Ling Suah [1] and Saw Lan Ong[2]

[1-2]*University of Science, School of Educational Studies, Malaysia*

ARTICLE INFO

Article History:
Received 11.11.2011
Received in revised form 29.02.2012
Accepted Tarih girmek için burayı tıklatın.
Available online 02.04.2012

ABSTRACT

The objectives of this study were to investigate the assessment practices of in-serviceteachers and to compare the assessment practices of teachers in different subject areas, teaching levels and teaching experience. Altogether 406 in-service teachers responded to the Teacher Assessment Practice Inventory. Rasch's model was used to analyse the characteristics of the assessment practices adopted by the teachers. Differential item functioning was performed to compare the assessment practices. In-service teachers were found to often use traditional types of assessment. The assessment practices differed between language teachers and science and mathematics teachers, primary school teachers and secondary school teachers and experienced teachers with inexperienced teachers.

© 2012 IOJES. All rights reserved

Keywords:
Assessment practice, Rasch's model, differential item functioning

Introduction

Assessment of student learning is an essential component of school activities. Research indicates that a sizable amount of classroom time is devoted to the assessment of student learning. Teachers spend between 10% to 50% of classroom time in assessment related activities (MacBeath & Galton, 2004; Stiggins, 2001).

Information from assessment is used for numerous purposes: to grade students, to group students, to diagnose student needs, improve students' motivation to learn, and to evaluate instruction (Brookhart, 1999). Assessing student performance is one of the most critical aspects of the job of a school teacher. Most of the assessment activities in the school are conducted by teachers. This underscores the need for a high level of assessment competency among in-service teachers.

The educational reform has called for the implementation of multiple sources of assessment information from the classroom instead of just relying on the summative one-time examination (Linn & Miller, 2005). The Malaysian Ministry of Education has responded to this assessment reform and drafted a new national assessment system for all public schools. The thrust of the change was to reduce reliance on the highly-centralized examination system to a system that integrates school-based assessment with the centralized examination. In anticipation of the reformation of the assessment system, the current assessment practices of in-service teachers need to be known so that appropriate action can be taken to improve the assessment skills of in-service teachers. As assessment practices of Malaysian teachers are not well explored, this study was carried out to identify the current assessment practices of in-service teachers in the northern states of Peninsular Malaysia. In addition, this study examined the differences in assessment practices between secondary and primary school teachers, language and science and mathematics teachers, and novice and experienced teachers.

[2] Corresponding author's address University of Science, School of Educational Studies, Penang, Malaysia
Telephone 604-6533240
Fax : 604-6572907
e-mail: ong_sawlan@yahoo.com; osl@usm.my

© 2012 International Online Journal of Educational Sciences (IOJES) is a publication of Educational Researches and Publications Association (ERPA)

Research Questions

Specifically, this study addressed the following research questions:

1. What are the common assessment practices of in-service teachers?

2. Are there any differences in teacher assessment practices between secondary and primary school teachers?

3. Are there any differences in teacher assessment practices between language and science and mathematics teachers?

4. Are there any differences in teacher assessment practices based on years of teaching experience?

Literature on Classroom Assessment

Classroom assessment serves many purposes for teachers, including grading, identification of student special needs, student motivation, and monitoring of instructional effectiveness (Ohlsen, 2007). The main purpose of classroom assessment, however, is to gather information about a student's learning (Abu Bakar Nordin, 1986; Airasian, 2001; Desforges, 1989; Jacobs & Chase, 1992; McMillan, 2008). Conducting classroom assessment is no simple task as it embraces a broad spectrum of activities which include constructing paper-and-pencil tests and performance measures, grading, interpreting test scores, communicating assessment results and using assessment results in decision-making.

When selecting a test format, teachers should be aware of and understand the strengths and weaknesses of the various assessment methods, and choose the one that best fits the different achievement targets (Stiggins, 1992). Only then can teachers use the appropriate assessment terminology and communication techniques to communicate the assessment results effectively to the target group (Stiggins, 1997). Teachers should be able to use the test scores appropriately and identify diagnostic information from the test results about instruction and student learning (Airasian, 2001). In the Malaysian education system, teachers are also expected to make decisions about students' educational placement, promotion, and graduation based on the assessment results.

According to Chang (1988), most teachers prefer to use tests and examinations to assess students' learning, especially English language teachers. Classroom teachers were shown to often use the paper-and-pencil tests (Abu Bakar Nordin, 1986; Airasian, 2001; Stiggins & Bridgeford, 1984), performance assessments, authentic assessments, and informal assessments such as observation and questioning to obtain information on student learning (Airasian, 2001; Stiggins & Bridgeford, 1984). In the paper-and-pencil test, the most commonly used item formats were the multiple choice and essay questions (Gullickson, 1993).

From a summary of the expectations of the assessment community towards school teachers, Schafer (1989) suggested eight areas of assessment skills that teachers need to develop. They are basic concepts and terminology of assessment; use of assessment; assessment planning and development; interpretation of assessment; feedback and grading; ethics of assessment; description of assessment results; and evaluation and improvement of assessment.

In 1990, the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) issued seven Standards for Teacher Competence in Educational Assessment of Students. The Standards specify that teachers should be skilled in choosing assessment methods; developing assessment methods; administering, scoring and interpreting assessment results; using assessment results for decision making; grading; communicating assessment results; and recognizing unethical assessment practices.

Stiggins (1999), however, asserts that these standards are not comprehensive enough to prepare teachers for the realities they will face in the classroom. Instead, he listed seven competencies: connecting assessment to clear purposes; clarifying achievement expectations; applying proper assessment methods; developing quality assessment exercises and scoring criteria and sampling appropriately; avoiding bias in assessment; communicating effectively about student achievement; and using assessment as an instructional intervention. Many of these were included in the Standards.

Teacher Assessment Practices

Studies focusing on classroom assessment showed that teacher assessment practices have been affected by subject areas (Bol, Stephenson, & O'Connell, 1998; Marso & Pigge, 1987, 1988; McMorris & Boothroyd, 1993; Zhang & Burry-Stock, 2003), school level (Bol, et al., 1998; Marso & Pigge, 1987, 1988; Mertler, 1998; Trepanier-Street, McNair, & Donegan, 2001; Zhang & Burry-Stock, 2003) and years of teaching experience (Bol, et al., 1998; Mertler, 1998). As expected, mathematics teachers tend to use more problem-solving items (Marso & Pigge, 1987, 1988) and calculation items (McMorris & Boothroyd, 1993). Marso and Pigge (1988) found that science and mathematics teachers relied more on paper-and-pencil tests rather than informal assessment procedures in contrast to the mathematics teachers in Bol et al.'s study (1998) who were not in favor of traditional assessment. In the case of item format, language teachers used more essay items to assess student learning (Marso & Pigge, 1987, 1988) while science teachers preferred multiple-choice items instead (McMorris & Boothroyd, 1993). Teachers of all subject areas commonly used paper-and-pencil tests (Zhang & Burry-Stock, 2003).

Several studies comparing primary school teachers with secondary school teachers found that primary school teachers frequently used alternative assessment or performance assessment (Bol, et al., 1998; Mertler, 1998; Zhang & Burry-Stock, 2003) and informal assessment in the form of observation and questions (Mertler, 1998). On the other hand, secondary school teachers used traditional types of assessment more often (Mertler, 1998) such as paper-and-pencil tests in the form of multiple-choice items (Mertler, 1998; Zhang & Burry-Stock, 2003), essays and problem type items (Marso & Pigge, 1987, 1988). They were also constructing items of high cognitive levels (Marso & Pigge, 1987, 1988).

In terms of years of teaching experience, there was no significant difference on the use of traditional assessments. The results on the use of alternative assessments were inconsistent. Teachers with less teaching experience in Mertler's study (1998) as well as the experienced teachers in Bol et al.'s study (1998) both reported using alternative methods of assessment more frequently.

Rasch Model

The Rasch model is in the family of the item response theory (IRT) models. This model describes the relationship between the probability of endorsing an item and the person's ability (Bejar, 1983). The Rasch model assumes that item difficulty is the only item characteristic affecting an individual's performance on an item (Baker & Kim, 2004). The Rasch model provides estimates of item difficulty, estimates of a person's ability and a standard error of measurement for each item. The item difficulty and person ability parameters are estimated jointly to produce estimates that are reported in the unit of "logit". In this study, the Rasch model was used to investigate and compare the assessment practices of in-service teachers.

Differential Item Functioning

Differential Item Functioning (DIF) refers to a psychometric difference in how an item functions for two groups. In other words, DIF refers to a difference in item performance between two comparable groups of people (Dorans & Holland, 1993). DIF occurs when people from different groups with equal knowledge exhibit different probabilities of endorsing on an item (Schumacker, 2005). The presence of DIF in a particular item indicates that individuals having the same level of ability, but belonging to different groups, do not share the same expected response to the item (Penfield & Camilli, 2007; Roussos & Stout, 2004). The Rasch model states that differential item performance is due to the difference of item difficulty between the groups understudied (Linacre & Wright, 1987).

In this study, the DIF analysis was used to compare the assessment practices of teachers from different subject areas, teaching levels and years of teaching experience. The response patterns between the two groups were compared to identify items that functioned differently.

Method

Instrumentation

The instrument used in this study was the Teacher Assessment Practices Inventory (TAPI) which was developed specifically for this research. The constructs were identified from literatures on teachers' assessment practice. The items formulated undergone content validation by school teachers and experts in educational measurement. It also satisfied unidimensionality when checked with Rasch's model analysis. The results indTAPI consists of 57 items that describe assessment practices. For each item, the respondents were asked to report their assessment practices on a 5-point rating scale ranging from "NOT USED AT ALL" to "HIGHLY USED". Demographic information concerning gender, school level, subject areas and years of teaching experience were also collected.

TAPI was developed based on the Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME & NEA, 1990), Stiggins' (1999) Competencies of Assessment and Schafer's (1989) Knowledge of Assessment. Altogether five constructs were identified to cover a broad range of assessment activities including test construction, types of assessment, use of assessment, grading and scoring, and communicating assessment results.

A summary of the constructs, subscales and number of items is shown in Table 1.

Table 1. Constructs and subscales of TAPI

Constructs	Subscales	Number of items
Constructing test	Test development	5
	Sources of constructing test	6
	Cognitive level	6
Types of assessment	Traditional assessment	6
	Alternative assessment	5
	Informal assessment	5
Use of assessment	Formative assessment	7
	Summative assessment	3
Grading & scoring	-	10
Communicating assessment results	-	4

Confirmatory Factor Analysis of TAPI

The Model TAPI is tested with CFA using *Robust Maximum Likelihood analysis*. The fit indices are as shown in Table 2 are satisfactory, where indices NFI*, CFI*, IFI*, GFI and AGFI exceeded 0.90. In addition, values for SRMR and RMSEA* are less than 0.05 and 0.08 respectively.

Table 2: CFA of the TAPI model

Sample	n	NFI*	CFI*	IFI*	GFI	AGFI	SRMR	RMSEA*
Overall	203	0.924	0.928	0.928	0.955	0.918	0.041	0.062
Validation	203	0.909	0.917	0.917	0.947	0.907	0.045	0.069

Sample

Altogether 406 in-service teachers from the northern states of Peninsular Malaysia responded to TAPI. Almost two-thirds (68%) of the teachers were females and 32% were males. Nearly half (47.3%) of the teachers were language teachers and 52.7% were teaching Science and Mathematics. There were 64.3% of them teaching at the secondary level while only 35.7% were teaching at the primary level. As for the teaching experience, 45.4% of the teachers have had more than ten years of teaching experience and 54.6% with less than ten years of teaching experience.

Data Collection

Data were collected during the month of October 2009. TAPI were distributed to the in-service teachers in the northern states of Kedah, Penang and Perak with the assistance of graduates of the University who are school teachers. The respondents answered TAPI during their free time.

Data Analysis

The computer program WINSTEPS version 3.66 that is based on the Rasch model was used to estimate the item parameters for the 57 items in TAPI. Rasch Model provides estimates of item difficulty which are reported in units of "logit". Item difficulties of the 57 items of TAPI were estimated to identify the assessment practices of the in-service teachers. The lower the value of item difficulty (in terms of logit), the higher is the type of assessment practised by the teachers. Conversely, the higher item indices indicated less use of the assessment practice by the teachers. The mean value of each assessment subscale was computed to reveal the endorsement level of each assessment category.

The DIF analysis performed was to compare the teachers' assessment practices according to subject areas taught, teaching levels and years of teaching experience. The DIF analysis identifies items that display psychometric differences which signify that the items are functioning differently for the two different groups matched by the measured construct. An item is flagged as DIF if the Welch t-value is greater than 1.96 or less than -1.96 at $p<0.05$. The DIF category suggested by ETS (Educational Testing Services) are large, if DIF contrast $\geq$ 0.64, moderate if 0.43 $\geq$ DIF contrast $\geq$ 0.64 and negligible for DIF contrast 0.43.

Findings

Constructing Test

When developing an assessment, the matching of assessment to instruction has the lowest item parameter index (-1.02 logit), which indicates that the in-service teachers placed great importance on alignment between assessment and their teaching. However, the highest item value for preparation of a table of specifications (0.35 logit) as shown in Table 2 implied that teachers seldom set up a table of specifications when constructing tests. Revising a test based on item analysis has a slightly below average item parameter value (-0.29 logit) which means the teachers item information to construct classroom tests.

Table 2. Item Parameter Estimates for Test Development

Items	Item parameter estimates (Logit)	Standard error
Matching with instruction	-1.02	0.09
Adequate content sampling	-0.87	0.08
Based on clearly defined course objectives	-0.73	0.08
Revises a test based on item analysis	-0.29	0.06
Uses a table of specifications	0.35	0.06

For developing of items according to Bloom's taxonomy of cognitive levels, Table 3 shows that questions for comprehension has the lowest item index (-0.59 logit) with almost the same value (-0.58 logit) for application levels. This shows that teachers are developing mostly test items which are either comprehension or application of contents that the students have learned. Item for synthesis level has the highest value (0.24 logit), which rarely appear in test items prepare by teachers. Unexpectedly, evaluation, the second highest cognitive level, has a slightly lower item value (-0.20 logit), which means teachers felt that they have prepared more items of this cognitive level.

Table 3. Item Parameter Estimates for Cognitive Level of items

Items	Item parameter estimates (Logit)	Standard error
Comprehension	-0.59	0.09
Application	-0.58	0.07
Knowledge	-0.40	0.08
Analysis	-0.32	0.07
Evaluation	-0.20	0.07
Synthesis	0.24	0.07

For sourcing test items, Table 4 shows that selecting questions from text books has the lowest item parameter value (-0.17 logit) followed by revision books (-0.14 logit) and public examinations (-0.12 logit). Using questions by department head has the highest value (0.65 logit) which shows that the teachers rarely obtained items from this source. The teachers were found to not construct their own questions frequently (0.43 logit) or use other teachers' questions (0.51).

Table 4. Item Parameter Estimates on Sources of Test Items

Items	Item parameter estimates (Logit)	Standard error
Text book	-0.17	0.07
Revision book	-0.14	0.07
Questions from public examination	-0.12	0.06
Construct own questions	0.43	0.06
Other teachers' questions	0.51	0.06
Questions from department head	0.65	0.05

Types of Assessment

Among the six traditional assessment item formats, multiple-choice questions has the lowest item parameter estimate (-0.15 logit) which indicates that this is the item format favored by the in-service teachers. Short answer questions (0.10 logit) and essay questions (0.24 logit) as shown in Table 5 are another two popular item format. Both the matching questions (0.93 logit) and true/false type of questions (0.90 logit) have high and almost comparable item parameter values which means the teachers seldom used these two types of items.

Among the performance assessments, homework was the most commonly used form of assessment as it has the lowest item parameter (-0.13 logit) as shown in Table 6. Project work has the highest item value (1.11 logit) which means it was rarely used. Similarly, both practical work and assignment were not well adopted by the teachers where both have item parameter estimates of 0.90 logit.

Table 5. Item Parameter Estimates of the Traditional Assessment Item Format

Items	Item parameter estimates (Logit)	Standard error
Multiple-choice questions	-0.15	0.06
Short answer questions	0.10	0.06
Essay questions	0.24	0.05
Fill in the blanks questions	0.50	0.05
True/false questions	0.90	0.05
Matching questions	0.93	0.05

Table 6. Item Parameter Estimates of the Alternative Assessment Techniques

Items	Item parameter estimates (Logit)	Standard error
Homework	-0.13	0.06
Practical work	0.90	0.05
Assignment	0.90	0.05
Portfolio	1.10	0.05
Project	1.11	0.05

In the case of informal assessment strategies, oral questioning has the lowest item estimate (-0.47 logit) followed closely by observations (-0.41 logit) as presented in Table 7. The results indicate in-service teachers frequently used these two types of informal assessments. The use of students' self ratings have the highest item value (0.65 logit) followed by interviews (0.53 logit) which means the teachers seldom used these two strategies.

Table 7. Item Parameter Estimates of the Informal Assessment Strategies

Items	Item parameter estimates (Logit)	Standard error
Oral questioning	-0.47	0.07
Observations	-0.41	0.07
Groupwork	0.35	0.06
Interviews	0.53	0.06
Student's self ratings	0.65	0.06

Uses of Assessment

In the use of assessment for formative purposes, providing feedback to students has the lowest item estimate (-0.83 logit) and a slightly higher value for identifying students' strengths and weaknesses (-0.73 logit) as is shown in Table 8. This results indicate that teachers have been giving feedback to students on their learning as well as helping them to identify their own strengths or weaknesses. The information, however, was not used by teachers to improve instruction in the classroom as the item estimate (-0.16 logit) is the highest.

Table 8. Item Parameter Estimates on Uses of Formative Assessment

Items	Item parameter estimates (Logit)	Standard error
Provide feedback to students	-0.83	0.08
Identify strengths & weaknesses of students	-0.73	0.08
Assign grades	-0.51	0.08
Improve students' motivation to learn	-0.47	0.08
Communicating academic expectations	-0.42	0.08
Grouping students	-0.23	0.07
Improve teachers' instruction	-0.16	0.08

Table 9 presents the results for the "Summative Use of Assessment". The use of assessment to determine students' grade has the lowest item estimate (-0.57 logit) followed by the measure of the students' achievement (-0.45 logit) and ranking of students (-0.30 logit). The item parameter estimates were all of negative values which imply that these practices are commonly adopted by teachers.

Table 9. Item Parameter Estimates on Uses of Summative Assessment

Items	Item parameter estimates (Logit)	Standard error
To determine a grade	-0.57	0.08
To measure a student's achievement	-0.45	0.09
To rank students	-0.30	0.07

Grading and Scoring. For grading and scoring of students' work as is given in Table 10, giving encouraging comments has the lowest item estimate (-0.41 logit) and is, thus, being practised frequently. The teachers also considered effort put in by the students when giving grades as it has the second lowest item estimate (-0.28 logit). Attendance, however, was often not taken into consideration in the calculation of grades with the highest item estimate obtained (0.26 logit). Neither were the teachers giving descriptive feedback often as the item estimate is the second highest (0.24 logit).

Table 10. Item Parameter Estimates on Grading and Scoring

Items	Item parameter estimates (Logit)	Standard error
Give encouraging comments	-0.41	0.07
Incorporate effort in the calculation of grades	-0.28	0.07
Use numerical score	-0.10	0.06
Incorporate class participation in the calculation of grades	-0.06	0.07
Descriptions of the extent to which goals were met	-0.03	0.07
Use letter grades	-0.01	0.06
Incorporate teamwork in the calculation of grades	0.07	0.06
Incorporate classroom behaviour in the calculation of grades	0.14	0.06
Provide descriptive feedback	0.24	0.07
Incorporate attendance in the calculation of grades	0.26	0.06

Communicating Assessment Results

Teachers frequently conveyed the assessment results to their students as reflected by the lowest item estimate (-0.55 logit) shown in Table 11. Communicating assessment results to the school administrator has the highest item difficulty (0.87 logit) followed closely by parents (0.64 logit). This means the teachers rarely reported the assessment results to them.

Table 11. Item Parameter Estimates on Communicating Assessment Results

Items	Item parameter estimates (Logit)	Standard error
Students	-0.55	0.07
Other educators	0.08	0.07
Parents	0.64	0.06
School's administrator	0.87	0.06

Differences in Teacher Assessment Practices Based on School Level

For this comparison, the DIF analysis was performed with the primary school teachers (N=145) constitute the focal group while the reference group is made up of secondary school teachers (N=261). There were 12 items identified as functioning differently between the primary and secondary school teachers as shown in Table 12. Secondary school teachers differ from primary teachers in developing tests based on the content of the subject ($t=2.97$, $p<.05$) and sourced test questions from the past-years' public examinations

(t=3.19, p<.05). In communicating test results, secondary teachers frequently provided descriptive feedback to the students (t=2.43, p<.05) while primary school teachers communicated test results to the parents (t=-2.33, p<.05). In the case of alternative assessment, secondary school teachers used more homework (t=2.75, p<.05) and coursework (t=2.31, p<.05) to assess student learning. They were also tend to provide opportunities for students to carry out self-assessments (t=2.29, p<.05).

Differences in Teacher Assessment Practices Based on School Level

For this comparison, the DIF analysis was performed with the primary school teachers (N=145) constitute the focal group while the reference group is made up of secondary school teachers (N=261). There were 12 items identified as exhibiting DIF as shown in Table 12 but only three items are moderate DIF while the rest are negligible. Secondary school teachers differ from primary teachers in developing tests based on the content of the subject (t=2.97, p<.05) and sourced test questions from the past-years' public examinations (t=3.19, p<.05). In communicating test results, secondary teachers frequently provided descriptive feedback to the students (t=2.43, p<.05) while primary school teachers communicated test results to the parents (t=-2.33, p<.05). In the case of alternative assessment, secondary school teachers used more homework (t=2.75, p<.05) and coursework (t=2.31, p<.05) to assess student learning. They were also tend to provide opportunities for students to carry out self-assessments (t=2.29, p<.05).

In the use of traditional assessment, primary school teachers used more filling in the blank questions (t=-3.19, p<.05), true/false questions (t=-3.47, p<.05), matching questions (t=-4.24, p<.05), oral questioning (t=-2.17, p<.05) and observation (t=-3.13, p<.05) to assess student learning as indicated in Table 12.

Table 12. DIF between Primary and Secondary School Teachers

Measure of Primary Teachers	Measure of Secondary Teachers	DIF Contrast (Logit)	Welch t-value*	DIF category	Items
-0.68	-1.22	0.54	2.97	moderate	Develop a test based on the teaching content
0.13	-0.27	0.40	3.19	negligible	Select test questions from public examinations
0.28	0.62	-0.34	-3.19	negligible	Fill in the blanks questions
0.66	1.03	-0.38	-3.47	negligible	True/false questions
0.64	1.09	-0.45	-4.24	moderate	Matching questions
0.07	-0.26	0.33	2.75	negligible	Homework
1.06	0.82	0.24	2.31	negligible	Coursework
-0.70	-0.35	-0.34	-2.17	negligible	Oral questioning
-0.74	-0.24	-0.50	-3.13	moderate	Observation
0.83	0.55	0.28	2.29	negligible	Self assessment by student
0.45	0.12	0.33	2.43	negligible	Provide descriptive feedback
0.44	0.74	-0.30	-2.33	negligible	Communicating assessment results to parents

*p<.05

Differences in Teacher Assessment Practices According to Subject Areas

When comparing language teachers with science and mathematics teachers, seven items were identified as DIF but only one item categorised as moderate DIF. The analysis was performed with the language teachers (N=192) as the focal group and the science and mathematics teachers (N=214) as the reference group. Science and mathematics teachers frequently selected test questions from textbooks or revision books (t=2.10, p<.05)), or questions from public examinations (t=3.20, p<.05). As expected due to the nature of the subject, Science and Mathematics teachers used more practical work (t=3.71, p<.05) and homework (t=3.09, p<.05) to assess student learning compared with language teachers as is shown in Table 13.

On the other hand, language teachers used more essay questions (t=-2.67, p<.05) than Science and Mathematics teachers. In the reporting of results, language teachers reported that they used more of letter grades (t=-3.93, p<.05) and numerical scores (t=-2.11, p<.05) when grading students' work.

Table 13. DIF between Language and Science & Mathematics Teachers

Measure of LanguageTeachers	Measure of Science & Maths Teachers	DIF Contrast (Logit)	Welch t-value*	DIF category	Items
0.01	-0.28	0.29	2.10	negligible	Select test questions from textbook or revision book
0.08	-0.31	0.40	3.20	negligible	Using questions from the public examination
0.08	0.36	-0.28	-2.67	negligible	Essay questions
1.09	0.73	0.35	3.71	negligible	Practical work
0.05	-0.31	0.37	3.09	negligible	Homework
-0.28	0.19	-0.46	-3.93	moderate	Using letter grades
-0.24	0.02	-0.27	-2.11	negligible	Using numerical scores

*p<0.05

Differences in Teacher Assessment Practices According to Years of Teaching Experience

The DIF analysis between teachers with more than ten years of teaching experience and teachers with less than ten years of teaching experience identified eight items functioning differentially between the two groups. They were all categorized as negligible DIF. For the analysis, the experienced teachers (N=184) made up the reference group while the less experienced teachers (N=222) were the focal group. As shown in the Table 14, teachers with less than 10 years of experience (t=3.21, p<.05) tended to use test questions prepared by other teachers when constructing a test.

With regards to the use of traditional, alternative and informal assessment techniques, there were also differences between the two groups of teachers. Experienced teachers used more true/false questions (t=2.26, p<05) while less experienced teachers used more matching questions. Teachers with less experience seemed to adopt the alternative assessment with the use of projects (t=3.60, p<.05), practical work (t=3.63, p<.05), portfolio (t=2.87, p<.05) and coursework (t=2.39, p<.05) in assessing students' learning. However, experienced teachers (t=-2.54, p<.05) used more of oral questioning compared with the less experienced teachers.

Table 14. Comparison of DIF Measure of Items Based on Years of Experience

Measure of Teachers with <10 years of Teaching Experience	Measure of Teachers with >10 years of Teaching Experience	DIF Contrast (Logit)	Welch t-value*	DIF category	Items
0.33	0.72	0.39	3.21	neligible	Using questions that other teachers have developed
0.80	1.03	0.23	2.26	neligible	True/false questions
0.78	1.10	0.31	3.12	neligible	Matching questions
0.94	1.31	0.38	3.60	neligible	Projects
0.74	1.09	0.35	3.63	neligible	Practical work
0.96	1.26	0.30	2.87	neligible	Portfolio
0.79	1.03	0.24	2.39	neligible	Coursework
-0.30	-0.68	-0.38	-2.54	neligible	Oral questioning

*p<0.05

Discussion

This study revealed that in-service teachers used more traditional types of assessment compared to alternative assessment. This may be attributed by the lack of knowledge and skills in alternative assessment during their teacher education program which resulted in their inability to put it into practice. This is especially obvious for teachers who have been teaching for more than 10 years. There is a need for more professional development programs on enhancing teachers' ability in carrying out alternative assessments.

Like the teachers in Gullickson's (1993) study, teachers in this study were found to depend very much on traditional assessment techniques such as multiple-choice questions, short answer questions and essay questions. This practice may be due to the influence of the public examinations in the Malaysian education system which are mostly in the form of multiple-choice, essays and short-answer questions type (Author et al, 2010). As the results of the public examinations are high-stakes and play an important role in determining the students' future, teachers assess students' learning according to the format of the public examinations to ensure that students are well prepared for the examinations and can succeed in these examinations.

When developing a test, teachers often did not prepare a table of specifications to help them in the planning of the number of items in each content area as well as determining the cognitive levels of the items. Ignoring this test development step means that they did not ensure the establishment of content validity of the test. In addition, they seldom constructed their own questions or revised the test items based on information obtained from the item analysis. One possible reason may be due to the lack of knowledge or skills required to carry out the analysis. In terms of feedback, teachers did provide feedback to students regarding their strengths and weaknesses of their learning.

The teachers in this study used different assessment practices according to their subject areas, school levels and years of teaching experience. These results are in tandem with those of Bol et al. (1998), Marso and Pigge (1987, 1988), McMorris and Boothroyd (1993), Zhang and Burry-Stock (2003), Mertler (1998) and Trepanier-Street, McNair, and Donegan (2001). Primary school teachers used more filling in the blanks questions, true/false questions, matching questions and portfolios to assess student learning but less of essay questions. Secondary school teachers used more summative assessments and scoring rubrics to determine the grades. In communicating test results, secondary school teachers provided descriptive feedback to the students themselves but primary school teachers often communicated the test results to the parents and school administrator. At the secondary level, students are more matured and could take appropriate actions

based on the teachers' inputs whereas students in the primary schools were not able to comprehend the meanings of the feedback given by the teachers.

The assessment practices between language teachers and science and mathematics teachers differed in several aspects. Language teachers used more essay questions but Science and Mathematics teachers used more practical work and homework to assess student learning. This is also indicated by Marso and Pigge (1988). The Science and Mathematics teachers used more of alternative assessments than the language teachers.

The teaching experience of teachers, too, had an effect on the assessment practices. The junior teachers who had less teaching experience used alternative assessment more frequently than the senior experienced teachers. This pattern was also seen in Mertler's study (1998). However, teachers with less experience were not able to construct their own test questions and resorted to using test questions from other teachers. This may be attributed by the lack of the necessary skills to develop good quality items and, hence, need professional development training in this area of assessment.

Conclusion

Teachers' assessment practices differed according to the school level, subject areas and also teaching experience. These results imply that teacher training programs for assessment cannot be of a standard type for all teachers. Assessment training needs to be diverse to cater to the different needs of different teachers. Since several differences were found between teachers at different levels of education (secondary and primary schools) and different subject areas (language and Science and Mathematics), wherever possible the content of the teacher training programs should be modified to cater to the needs of the level at which the pre-service teachers will be teaching in the future. Teacher training programs need to address the actual needs of school teachers; only then can the teachers be considered to have been adequately prepared to assess students' performance. In addition, more emphasis on techniques of alternative assessment should be given to teachers to ensure accurate and effective assessment.

References

Abu Bakar Nordin (1986). *Asas penilaian pendidikan*. Petaling Jaya: Longman Malaysia Sdn Bhd.

Airasian, P. W. (2001). *Classroom assessment: Concepts and applications* (4th ed.). New York: McGraw-Hill Higher Education.

American Federation Of Teachers, National Council On Measurement In Education, & National Education Association (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues & Practice, 9*(4), 30-32.

Author et al (2010) [details removed for peer review]

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Deeker, Inc.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 1-23). Vancouver: Educational Research Institute of British Columbia.

Bol, L., Stephenson, P. L., & O'Connell, A. A. (1998). Influence of experience, grade level and subject area on teachers' assessment practices. *Journal of Educational Research, 91*(6), 323-330.

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

Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*: Washington DC ERIC Clearinghouse On Higher Education And Office Educational Research And Improvement.

Chang, S. F. (1988). *Teachers' assessment practices: Assessing phase II pupils' progress in KBSR English.* Unpublished master's thesis, Universiti Malaya, Petaling Jaya.

Desforges, C. (1989). *Testing and assessment*. London: Cassell Education Limited.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mental-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). New Jersey: Lawrence Erlbaum Associates.

Gullickson, A. R. (1993). Matching measurement instruction to classroom-based evaluation: Perceived discrepancies, needs, and challenges. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 1-25). Lincoln, NE: Buros Institute of Mental Measurement, University of Nebraska-Lincoln.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

Jacobs, L. C., & Chase, C. I. (1992). *Developing and using tests effectively*. San Francisco: Jossey-Bass Publishers.

Linacre, J., & Wright, B. D. (1987). Item bias:Mantel Haenszel and the Rasch Model Retrieved 20 November, 2009, from http://www.rasch.org/memo39.pdf

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.). New Jersey: Pearson Education.

MacBeath, F., & Galton, M. (2004). A life in secondary teaching: Finding time for learning Retrieved 23 March, 2009, from http://www.data.teachers.org.uk/resources/pdf/74262-MacBeath.pdf

Marso, R. N., & Pigge, F. L. (1987, October). *Teacher-made tests and testing: Classroom resources, guidelines, and practices*. Paper presented at the Annual Meeting Of The Midwestern Educational Research Association, Chicago.

Marso, R. N., & Pigge, F. L. (1988, April). *An analysis of teacher-made tests: Testing practices, cognitive demands, and item construction errors*. Paper presented at the Annual Meeting Of The National Council On Measurement In Education, New Orleans, Louisiana

McMillan, J. H. (2008). *Assessment essentials for standard-based education* (2nd ed.). California: Corwin Press.

McMorris, R., & Boothroyd, R. (1993). Tests that teachers build: An analysis of classroom tests in Science and Mathematics. *Applied Measurement in Education, 6*(4), 321-342.

Mertler, C. A. (1998, October). *Classroom assessment: Practices of Ohio teachers*. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago.

Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). New Jersey: Pearson Education.

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

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning form to improve the interpretation of DIF in polytomous items. *Applied Measurement in Education, 22*(1), 61-78.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 26, pp. 126-167). New York: Elsevier.

Reckase, M. (1979). Unifactor latent models applied to multi-factor tests: Results and implication. *Journal of Education Statistics, 4*(4), 207-230.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis: Detecting DIF item and testing DIF hypotheses. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-115). Thousand Oaks: Sage.

Schafer, W. D. (1989). *Assessment essentials in professional education of teachers*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Schumacker, R. E. (2005). Test bias and Differential Item Functioning Retrieved 22 May, 2009, from www.appliedmeasurementassociates/testbias&dif.pdf

Smith, A. B., Wright, E. P., Rush, R., Stark, D. P., Velikova, G., & Selby, P. J. (2006). Rasch analysis of the dimensional structure of the hospital anxiety and depression scale. *Psycho-Oncology, 15*(9), 817-827.

Stiggins, R. J. (1992). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice, 11*(2), 35-39.

Stiggins, R. J. (1997). *Student-centered classroom assessment*. New York: Merrill Publishing.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23-27.

Stiggins, R. J. (2001). The principals' leadership role in assessment. *NASSP Bulletin, 85*(13), 13-26.

Stiggins, R. J., & Bridgeford, N. J. (1984). *The use of performance assessment in the classroom*. Portland: Northwest Regional Educational Lab.

Trepanier-Street, M. L., McNair, S., & Donegan, M. M. (2001). The views of teachers on assessment: A comparison of lower and upper elementary teachers. *Journal of Research in Childhood Education, 15*(2), 234-241.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8(3), 370.* Retrieved June 23, 2009, from http://www.rasch.org/rmt/rmt83.htm

Zhang, Z., & Burry-Stock, J. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323-342.

Appendix

Inventori Amalan Pentaksiran Guru (IAPG)

Inventori ini bertujuan memperoleh maklumat tentang amalan-amalan pentaksiran guru dalam bilik darjah.

Arahan: Untuk pernyataan di bawah, sila beri respons anda dengan membulatkan nombor yang sesuai.

Untuk setiap pernyataan, sila gunakan skala berikut:

1- Tiada 2- Jarang-jarang 3- Selalu 4- Sangat kerap

	(A) Pembinaan Ujian	Tiada	Jarang-jarang	Selalu	Sangat kerap
1.	Semasa anda membina ujian, berapa kerapkah anda.....				
	(a) menggunakan Jadual Penentuan Ujian (JPU)	1	2	3	4
	(b) merujuk kepada objektif pembelajaran	1	2	3	4
	(c) merujuk kepada kandungan pengajaran & pembelajaran	1	2	3	4
	(d) merujuk kepada sukatan pelajaran	1	2	3	4
	(e) menentukan bilangan item mengikut pemberatan isi kandungan yang diajar	1	2	3	4
2.	Berapa kerapkah anda mengubahsuaikan soalan daripada sumber- berikut untuk dijadikan soalan ujian anda?				
	(a) soalan daripada buku teks atau buku rujukan	1	2	3	4
	(b) buku ulang kaji	1	2	3	4
	(c) soalan ujian yang dibina oleh guru lain	1	2	3	4
	(d) soalan ujian yang diberi oleh ketua panitia	1	2	3	4
	(e) soalan daripada kertas peperiksaan awam	1	2	3	4
3.	Berapa kerapkah anda membina soalan ujian dengan aras kognitif berikut?				
	(a) Pengetahuan, iaitu mengingati fakta dan maklumat yang dipelajari	1	2	3	4
	(b) Pemahaman, iaitu memahami isi kandungan yang dipelajari	1	2	3	4
	(c) Aplikasi, iaitu mengaplikasi perkara yang dipelajari dalam situasi baru	1	2	3	4
	(d) Analisis, iaitu menganalisis isi kandungan yang dipelajari	1	2	3	4
	(e) Sintesis, iaitu mengsintesis maklumat yang dipelajari menjadi bentuk baru	1	2	3	4
	(f) Penilaian, iaitu membuat penilaian terhadap perkara yang dipelajari	1	2	3	4

	(B) Kaedah Pentaksiran	Tiada	Jarang-jarang	Selalu	Sangat kerap
1.	Apabila membina suatu ujian bertulis, berapa kerapkah anda menggunakan bentuk-bentuk pentaksiran berikut?				
	(a) soalan objektif aneka pilihan	1	2	3	4
	(b) soalan esei	1	2	3	4
	(c) soalan mengisi tempat kosong	1	2	3	4
	(d) soalan jawapan pendek	1	2	3	4
	(e) soalan betul/salah	1	2	3	4
	(f) soalan pemadanan	1	2	3	4
2.	Dalam menilai pelajar anda, berapa kerapkah anda menggunakan jenis-jenis penilaian berikut?				
	(a) projek	1	2	3	4
	(b) kerja amali	1	2	3	4
	(c) portfolio	1	2	3	4
	(d) projek kerja kumpulan	1	2	3	4
	(e) kerja kursus	1	2	3	4

3.	Berapa kerapkah anda menggunakan strategi-strategi di bawah untuk menilai pelajar anda?				
	(a) Menyoal pelajar secara lisan	1	2	3	4
	(b) Membuat pemerhatian terhadap pelajar	1	2	3	4
	(c) kerja rumah	1	2	3	4
	(d) latihan bertulis di bilik darjah	1	2	3	4
	(e) Pelajar membuat penilaian kendiri	1	2	3	4
	(f) Mengadakan temu bual dengan pelajar	1	2	3	4
(C). Penggunaan Hasil Pentaksiran		Tiada	Jarang-jarang	Selalu	Sangat kerap
1.	Berapa kerapkah anda gunakan hasil penilaian untuk tujuan berikut?				
	(a) mengenal pasti kelemahan pelajar	1	2	3	4
	(b) memotivasikan pelajar	1	2	3	4
	(c) memberi maklum balas kepada pelajar	1	2	3	4
	(d) memperbaiki pengajaran anda	1	2	3	4
	(e) mengetahui kemajuan pelajar	1	2	3	4
2.	Berapa kerapkah anda gunakan hasil penilaian untuk tujuan berikut:				
	(a) mengukur pencapaian pelajar	1	2	3	4
	(b) menentukan gred pelajar	1	2	3	4
	(c) mengumpul pelajar mengikut pencapaian	1	2	3	4
	(d) membanding pencapaian akademik di kalangan pelajar	1	2	3	4
(D) Penskoran & Penggredan		Tiada	Jarang-jarang	Selalu	Sangat kerap
1.	Apabila memeriksa hasil kerja pelajar, berapa kerapkah anda melakukan amalan berikut?				
	(a) menulis gred abjad seperti A, B, C, dsb.	1	2	3	4
	(b) menulis skor angka	1	2	3	4
	(c) memberikan maklum balas berbentuk deskriptif	1	2	3	4
	(d) memaklumkan sejauh manakah pelajar mencapai sasaran pembelajaran.	1	2	3	4
2.	Semasa anda menentukan gred pencapaian pelajar, berapa kerapkah anda mengambil kira perkara-perkara berikut?				
	(a) tingkah laku pelajar dalam bilik darjah	1	2	3	4
	(b) daya usaha pelajar	1	2	3	4
	(c) kehadiran	1	2	3	4
	(d) kerjasama kumpulan	1	2	3	4
	(e) penyertaan dalam kelas	1	2	3	4
(E) Maklum Balas Hasil Penilaian		Tiada	Jarang-jarang	Selalu	Sangat kerap
1.	Berapa kerapkah anda membincangkan kemajuan atau kelemahan pelajar dengan pihak-pihak berikut?				
	(a) pelajar	1	2	3	4
	(b) ibu bapa	1	2	3	4
	(c) guru-guru lain	1	2	3	4
	(d) pentadbir sekolah	1	2	3	4
2.	Berapa kerapkah anda mengamalkan perkara-perkara berikut?				
	(a) memberitahu pelajar secara lisan kesilapan pelajar yang telah dikesan melalui latihan mereka	1	2	3	4
	(b) memberi komen bertulis dalam latihan pelajar	1	2	3	4
	(c) memberi komen bertulis dalam laporan kemajuan pelajar	1	2	3	4