Pertanika J. Soc. Sci. & Hum. 7(1): 21 - 29 (1999)

ISSN: 0128-7702
© Universiti Putra Malaysia Press

The Perceptions of University Instructors on Their Testing Practices: A Case Study

MOHAMAD SAHARI
Kulliyah of Education
International Islamic University Malaysia
Jalan Gombak, 53100 Kuala Lumpur

Keywords: Testing, assessment, instructor-made-test, achievement test, higher education

ABSTRAK

Keupayaan pendidik dalam aspek pengujian dan penilaian telah diperakukan sebagai satu daripada petunjuk tingkat keberkesanan pengajaran. Yang demikian itu telah merangsang pelaksanaan penyelidikan ini yang secara khususnya disorotkan kepada amalan pengujian dan penilaian di peringkat pengajian tinggi. Kajian ini dilakukan dengan tujuan meninjau penggunaan prosedur penilaian yang sah dan boleh dipercayai dengan berdasarkan laporan para pensyarah di sebuah universiti awam. Sampel kajian terdiri daripada 105 tenaga akademik yang dipilih secara sistematis. Untuk merekodkan laporan pengujian dan penilaian yang diamalkan oleh mereka, penyelidik menggunakan satu senarai semak tentang prosedur membina soalan ujian akhir semester dan prosedur pemarkahan. Analisis deskriptif dan regresi logistik digunakan untuk menemukan jawapan terhadap masalah penyelidikan. Didapati sebahagian besar daripada responden melaporkan bahawa mereka mengamalkan langkah-langkah yang sejajar dengan prosedur pengukuran dan penilaian yang sah dan boleh dipercayai, terutamanya daripada segi pemarkahan. Juga ditemukan bahawa amalan pengujian yang sah adalah lebih ketara dan signifikan di kalangan pensyarah dalam bidang pendidikan, psikologi dan sains kemanusiaan berbanding dengan pensyarah dalam bidang lain. Kajian ini secara sederhana sekali membentangkan implikasi penemuan kepada amalan pendidikan, khususnya amalan pendidikan dan pensijilan tenaga akademik di institusi pengajian tinggi.

ABSTRACT

Instructors' ability and skills in creating valid and credible tests are the prime concerns that prompted this study. The focus of the study was to investigate prevalent testing practices among lecturers in higher learning institutions. Specifically it examined the procedures followed by university instructors in the constructions of valid and reliable tests. The study sampled 105 respondents systematically selected from a pool of university instructors. To obrain responses from the respondents regarding their testing practices, a checklist containing relevant items on the creation of final examination questions and their marking schemes was constructed. Data were analysed using descriptive statistics and logistic regressions. It was discovered that a majority of the respondents had applied proper procedures in creating valid and reliable tests, especially with respect to marking examination scripts. It appeared that instructors from the faculties of education, psychology and human sciences demonstrated adherence to these standards and proper procedures more significantly than those from other faculties. From the findings, the study synthesized implications on educational practices, in particular the certification of testing competency among instructors of higher learning institutions.

INTRODUCTION

Since much of college and university learning is assessed through instructor-made achievement tests, it would be useful to know how instructors develop and use them. However, research on student assessment by and large tends to focus

on the testing practices of the elementary and secondary school teachers. The results of research in this area have painted a somewhat disturbing picture. Despite their importance as a means of helping students to learn, elementary and secondary school teachers are generally poorly prepared to appropriately apply testing concepts, processes, and procedures.

Over the last few decades, much attention has been given to the importance of teachers' assessment to the degree that it has become an indicator of teacher effectiveness (Daniel & King, 1998). Shapiro (1995) pointed out that to satisfy the requirements set by the American National Board for Professional Teaching Standards, teachers must have adequate knowledge of testing and measurement procedures. On an equal note, Popham (1995) aptly stressed that public perceptions of educational effectiveness and teacher evaluation rely partly on teachers' assessment practices. One justification for this attention to teachers' testing and measurement competency is that proper assessment allows teachers to diagnose students' strengths and weaknesses, monitor their students' progress, assign grades, and determine their own instructional effectiveness (Popham, 1995). Furthermore, assessment has become an integral component of teaching in which a substantial portion of teachers' time is spent on assessing students' learning (Stiggins and Conklin, 1988; Stiggins, 1991). Stiggins (1991) estimated that teachers spend about one-half of their time on preparing and completing classroom tests.

Teachers' lack of knowledge of testing and measurement, which obviously undermines the quality of their assessment of students' achievement, is one dominant recurring finding yielded from previous works (Daniel & King, 1998; Goslin, 1967; Gullickson, 1984; Schafer & Lissitz, 1987; Noll, 1955; Roeder, 1972). Toward graduation, a majority of teacher certification programs do not require any course at all in educational testing and measurement (Goslin, 1967; Noll, 1955; Roeder, 1972; Wise, Lukin, & Roos, 1991), and this policy speaks volume for the assessment pitfall. Earlier Noll (1955) found that a course in testing and measurement was not a graduation requirement in public universities, and ironically two decades later, it was still not made a requirement in more than half of the teacher preparation programs (Roeder, 1972). It is, therefore, not surprising to find that almost one-half of the teachers surveyed by Roeder (1972) perceived that they had had inadequate training in assessment. Within the local context, only two out of six institutions of higher learning offer a course in measurement and evaluation to pre-service secondary school teachers (Ministry of Education, 1999).

Because of inadequate formal training in assessment, teachers failed to fully capitalize on its usefulness in helping students to learn; rather the literature suggests serious flaws in assessment practices and the prevalence of misuse. Popham (1995) claimed that although teachers were concerned about the test they had constructed, they had failed to develop valid classroom measures, execute reliable scoring practices, and fairly interpret test results. Previous works related to testing practices also found that teachers tested students' content-mastery based on half-baked instructions, assessed merely students' recall ability, awarded zeros for incomplete answers, used inappropriate tests, did not inform students about the contents of the tests, and were unable to communicate the results effectively to their students (Canandy & Hotchkiss, 1989; Daniel & King, 1998; Gullickson, 1985; Hills, 1991). In most likelihood, teachers rely on their own fossilized trial-and-error experiences on testing, instead of applying deep-seated knowledge on assessment, to make high-stakes decisions about teaching and learning. This inadequacy may create damaging effects in teachers' decision-making about students' achievement, placement, promotion, remedial, and retention.

It is of particular importance to find out whether similar gloomy scenario prevails in the testing practices at higher learning institutions. Unfortunately, in this respect, not much evidence about university instructors' testing practices has been documented. Of the scanty literature on the testing practices at institutions of higher learning, a somewhat similar pattern of testing practices is observed. For example, instructors were found to face difficulty in constructing items testing for higher-order thinking (Diamond, 1998). Shifflett, Phibbs, and Sage (1997) found that there is a distinguishable discrepancy between students' and instructors' attitudes regarding the fairness of tests. In addition, Freeman and Lewis (1998) reiterated

the importance of test blueprints in developing and validating instructor-made tests. More recently, Palomba and Banta (1999) reemphasized the fact that "students need to know the overall purpose of the testing and how the information will be used" (pp. 160-161).

The purpose of the present study was to explore the use of testing and measurement concepts, principles and procedures by university instructors. The study addressed the following research questions: (1) Do university instructors apply content-valid procedures in developing end-of-semester test? (2) Do they apply reliable scoring procedures? (3) Is there any relationship between instructors' experiences in assessing students' achievement and their application of valid and reliable testing procedures? While data on the first two questions would provide understanding about assessment as practiced at university level, the information elicited from the third question may shed light on the differing effects of knowledge in testing and measurement between groups of faculty members. Overall, the study concerned two issues related to instructor-made achievement tests, first, the development of content-valid test, and second, the reliability of the scoring procedure.

More than anything, the worth of a classroom achievement test rests on the content-related validity of the test (Anastasi, 1988; Oosterhof, 1990). By definition, content validity is a gauge built into a test in which the validity "is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn" (Messick, 1992, p. 1489). Typically, it serves as a yardstick to estimate how well a test fulfills whatever claim(s) it makes; it documents the extent to which the test measures what it purports to measure. This conception of validity derived a set of testing principles that require an instructor to, (1) define the purpose of the test, (2) describe the process of developing the test, (3) provide evidence that the test meets its intended purpose, and (4) keep all test takers equally informed about how much the test covers and the types of question format (Code of Fair Testing Practices in Education, 1988). Thus, to establish the validity of an achievement test, in particular a high-stakes final test, instructors should effectuate these principles into workable procedures.

Contemporary textbook writers suggest that content validity should be built into a test even before the test is constructed. One procedure to create a content-valid test is by drawing up a *table of specifications* (Anastasi, 1988), the contents of which comprise the topics and instructional objectives to be tested, and the sampling distribution of items that reflects the relative importance of individual topics and objectives. Another procedure involves a listing of performance objectives (Oosterhof, 1990). Gagne, Briggs and Wager (1988) propounded that each performance objective calls for identification of the specific capability, situation, object, action, and condition under which a student is to be tested. These procedures provide room for an expert–a person who is well-versed in the content knowledge to be tested–to judge the extent to which the test adequately samples the relevant contents, that is, whether the test covers a representative sample of concepts and skills it claims to measure. In effect, it is the opinion of an expert that validates the quality of a particular test.

The quality of a test is influenced by its reliability whose importance comes second after validity. A test is reliable to the extent that it measures something consistently. With respect to an achievement test, reliability manifests itself in producing consistent scores; hence systematic scoring procedures are a prerequisite to validity. Depending upon the purpose of the test, answers to essay items can be scored using either holistic or analytical procedure (Coffman, 1971); each has its own merits in educational assessment. While analytical procedure demands the preparation of a scoring plan that specifies attributes to be evaluated and awarded with points (Oosterhof, 1990), holistic scoring focuses on evaluating the overall quality of particular answers. Of these two techniques, holistic scoring is less reliable unless the procedure uses a second reader to independently examine the answers (Gronlund, 1985).

METHODOLOGY

Participants

The population of this study comprised instructors at an international university in Malaysia. Using a sampling frame consisting of 433 lecturers, associate professors, and professors, a systematic sampling procedure (Scheaffer,

Mendelhall, & Ott, 1996) was applied in order to draw a probability sample of 135 instructors. The sample comprised experienced instructors; the majority (90%) indicated familiarity with the university examination policy; 63% had been a course coordinator (course leader); 42.5% had attended at least one inservice educational development program which included a three-hour session on testing and measurement. In terms of demographic characteristics, 57.1% of the participants were male; 53.3% were Ph.D. holders; 37.5% specialized in areas related to psychology and education, the two areas of study closely associated with testing and measurement.

It should be noted that the sample size represents 24% of the total number of population under study. The background data indicated that there were no serious departures of the sample characteristics from the population. In particular, the distribution of participants according to demographic characteristics is comparable with that of the population parameters (Management Services Division, 1999).

Instrument

To identify how the participants practice assessment in their previous end-of-semester tests, the researcher used a checklist for test development and scoring procedures. Besides seeking demographic data, the checklist attempted to elicit information about whether or not the instructors applied valid and reliable procedures in developing and scoring the tests. More precisely, one section listed 12 items related to procedures used to develop a content-valid test, and another section contained eight items on scoring technique; each item required the participants to record whether they had applied the procedure. These 20 items were adapted from Oosterhof's (1990) criteria for evaluating achievement test. To estimate the reliability of the data, a test-retest correlation procedure was applied on the responses collected from a sample of 20 instructors who were not included in the study. Cramer's V coefficient correlation formula indicates that the reliability for each item ranges between .577 and 1.00 (Table 1 and Table 2).

To administer the instrument, eight graduate students undergoing a research methodology course visited the 135 randomly identified faculty members. They approached the participants individually to explain the purpose of the study, emphasize the principles of anonymity and confidentiality, and to solicit their cooperation. Each participant was allowed to self-administer the instrument and return it immediately to the assistant researcher. It was found that 30 of the instruments were either incomplete or could not be used at all, resulting with 77.8% rate of response ($\underline{N}$ = 105).

The sample size of 105 was adequate in terms of providing dependable estimates of the population characteristics. Using Scheaffer et al., (1996) approach to sample size determination, the error of estimation for the present study was ± 8.4%. In addition, a *post hoc* power analysis was conducted using Kraemer and Thiemann (1987) procedure. The purpose of the analysis was to identify the minimum level of statistical power, given the completed responses, at .05 alpha level and an effect size of $\underline{C}$ = .151; it is the smallest effect that would be of practical significance. The analysis detected the minimum power of 33.4% to correctly reject a no-difference hypothesis. This level of power corresponds to statistical power for most studies in human and social sciences (Sedlemeyer & Gigerenzer, 1989).

Analysis

One interest of the study was to establish a relationship between instructors' assessment practices and their demographic characteristics in terms of the experience that they had had in testing–their participation in the staff educational development program, experience as a course leader, areas of studies, and age. To address this concern, a logistic regression technique was applied on each of the three fundamental practices in assessing students' achievement; the three criterion variables being the participants' use of table of specifications, performance objectives, and scoring plan. Each of criterion variable measures a dichotomous outcome, which is whether the participant had applied the assessment procedure, as they perceived it. The predictors comprised the elements of the instructors' experiences in student assessment. For practical reasons, instructors' areas of studies were collapsed into three categories, (1) psychology and education, (2) other specialization in human sciences, and (3) other discipline of knowledge, which consisted of law, economics and management, and engineering and technology.

RESULTS

Application of Valid Procedures in the Test Development

Descriptive analysis of the data produced several noteworthy results (Table 1). First, more than two-thirds of the participants reported that they specified the intended content of the test (71.2%), and they made it a point to convey the information to the students (80.8%). Second, the data indicated that a large majority of the instructors attempted to build a validation mechanism, a test blueprint into their test construction. The results showed that while 40 participants (38.5%) perceived that they had created table of specifications, 62 of them (59.6%) reported that they had defined the performance objectives. However, it should be noted that information regarding the nature and quality of the instructors' test blueprints is not available, and therefore the results should be interpreted cautiously.

Third, in terms of the construction of test items, most participants reported the using valid procedures. The data showed that the instructors ranked the relative importance of the contents to be tested (60.6%), determined the number of items according to the relative importance of the contents (55.8%), estimated the amount of time spent on each item (76%), and assessed the difficulty level of each item (65.4%). Finally, the application of procedures to validate the test was done by a limited number of instructors. Of participants who reported that they had used performance objectives as their test blueprints, less than one-half (47.1%) compared the test questions that they had constructed with the specified performance objectives. Also, only 17 instructors (13.5%) reported that they sought expert opinion–the opinion of other instructors–to evaluate the test blueprint. Simply stated, although the majority of the faculty members applied valid procedures to develop end-of-semester test, they failed to establish the validity.

Application of Reliable Scoring Procedures

Of the 105 participants, 83 (79%) reported that they used analytical scoring procedures to evaluate students' performance on tests (Table 2). With the exception of awarding points for presentation style and use of language, more than two-thirds of the 83 participants applied procedures that contributes to reliable scoring. A majority of the participants indicated that they

TABLE 1
Application of valid procedures in the development of end-of-semester test as reported by the instructors (N = 105)

Procedures	%	r*
Specify the intended contents of the end-of-semester test.	71.2	1.000
Consider the students' inputs in deciding the contents of the test.	39.4	.903
Define the performance objectives to be tested.	59.6	.577
Rank the relative importance of the contents.	60.6	.601
Determine the number of questions according to the relative importance of the topics to be tested.	55.8	.787
Create a table of specifications to plan for the test.	38.5	.816
Have other instructors evaluate the table of test specifications.	13.5	.793
Inform the students about the format of question.	80.8	.688
Ascertain that the reading skills required by each question below that of students' ability.	31.7	.688
Assess the level of difficulty for each item.	65.4	.724
Estimate the amount of time spent on each question.	76.0	1.000
Compare the test questions to the specified performance objectives	47.1	.612

* Reliability Index.

had developed a scoring plan when the test was written (77.1%). Accordingly, they specified the total number of points each item was worthy of (81.9%), listed the attributes to be evaluated for each question (70.0%), determined that the points associated with each question were proportional to the relative importance of the content being tested (75.9%), and specified guidelines for awarding points to student's answers (79.5%). Interestingly almost all instructors (96.4%) sought the judgment of their colleagues on the accuracy of the scoring plan. Thus, the data indicated that relatively more instructors had used procedures related to reliability than they did the procedures for developing valid tests.

Although the participants generally had adhered with the requirements of reliable scoring, the data yielded one discomforting finding. The results indicated that while only 47 of the 83 instructors awarded points for presentation style and use of language, almost all of them (95.2%) penalized students for grammatical and spelling errors. This finding suggests the likelihood of lecturers awarding points and penalizing students' achievement on the basis of their language skills, rather than the mastery of content-knowledge. Such a practice, if continuously exercised, would underscore the reliability, and thereby the validity of student assessment.

Assessment Practices and Instructors' Experiences in Testing and Measurement

Table 3 summarizes the frequency and percentage distribution of the instructors' application of three important assessment procedures, which were whether they had constructed (1) table of test specifications, (2) performance objectives, and (3) scoring plans across categories of the participants' experiences in testing and measurement. This study included three elements of instructors' experience in student assessment; the elements were their participation in staff educational development program (SEDP), experience as a course leader, and areas of study. The variable age was used as a covariate. To explore the likelihood of a relationship between instructors' assessment practices and pertinent attributes of their experiences, the study analyzed three separate logistic regression models, each of which tested one of the three assessment procedures.

The results of logistic regression analyses indicated that only the instructor's assessment practice in terms of creating table of specifications was reliably related to their experiences in testing and measurement. Using the likelihood ratio to examine the overall relationship, the value of 2lnLH containing only a constant was 102.76 while that of the full model was 85.72. Thus, an addition of the predictors into the model yielded a reduction in

TABLE 2
Application of analytical scoring procedures in the end-of-semester test as reported by the instructors (N = 83)

Procedures	%	r*
Develop a scoring plan when the test question is written.	77.1	1.000
List the attributes to be evaluated for each question.	70.0	1.000
Specify the total number of points each item is worth.	81.9	.688
Determine that the points associated with each question are proportional to the relative importance of the content being tested.	75.9	.787
Specify guidelines for awarding points to student's answers.	79.5	.577
Seek other instructor judgement on the accuracy of the scoring plan.	96.4	.630
Award points for presentation style and use of language.	56.6	.798
Penalize students for spelling and grammatical error	95.2	.688

* Reliability index.

TABLE 3
Application of assessment procedures according to participants' attributes of testing experiences

	Table of Test specifications		Performance objectives		Scoring plan		
	YES	NO	YES	NO	YES	NO	n
Educational Development Program							
Attended	46.6	53.4	67.2	32.8	70.7	29.3	58
Never Attended	32.4	67.6	50.0	50.0	67.6	32.4	34
Course Leadership							
Yes	32.8	67.2	59.0	41.0	66.1	33.9	61
No	47.2	52.8	61.1	38.9	72.2	27.8	36
Area of Studies							
Psychology and Education	58.3	41.7	75.0	25.0	72.2	27.8	36
Human Sciences	52.6	47.4	63.2	36.8	57.9	42.1	19
Other Disciplines	17.5	82.5	40.0	60.0	68.3	31.7	40
N	38	57	37	60	38	54	

the lack of fit, resulting in a statistically overall relationship; $\chi 2(df = 5) = 17.05$, $p = .004$.

The logit equation for the model was $g = - 2.645 + .023(\text{Age}) + (-.022)(\text{SEDP}) + (.509)(\text{Leader}) + (1.773)(\text{Psychology and Education}) + (7.401)(\text{Human Sciences})$. However, of the four independent variables, only the instructors' area of study was found to be significantly associated with the likelihood of they creating a table of specifications to plan for the end-of-semester test. Partialling out the effects of other predictors, the odds of constructing a table of test specifications by those who specialized in psychology and education were about six times the odds for those in other disciplines of knowledge. Also, the odds of applying the assessment procedure among those who were in other area of human sciences were about seven times the odds of those who specialized in other disciplines of knowledge. Therefore, it can be inferred that it is more likely for an instructor who specialized in psychology, education, and other areas of human sciences to create a table of test specifications compared to those in law, economics and management, and engineering and technology.

CONCLUSION

The study found that a majority of the university instructors applied content-valid procedures in developing end-of-semester tests, and they adhered to reliable scoring procedures. By and large the participants conformed to the standards

TABLE 4
Summaries of the results of logisitic regression
on the instructors' application of table of test specifications

Elements of Experiences in Testing and Measurement	Effect	S.E.	p-value	Estim
Age	.023	.027	.402	1.02
SEDP	-.002	.573	.969	.98
Course Leader	.509	.557	.361	1.66
Area of Sudies				
Psychology and Education	1.773	.656	.006*	5.89
Other Human Sciences	2.002	.831	.016*	7.40

* Significant at alpha = .05.

that were deemed important, for example by the Code of Fair Testing Practices in Education (1988). In particular, the instructors generally defined the purpose of end-of-semester tests, and they did inform their students about the contents to be covered in the tests and the types of question format. The study also found that although university instructors made an effort to set up validation mechanism by using table of test specifications and performance objectives, they failed to substantiate the validity of their tests. Finally, more of the instructors of psychology, education, and other human sciences reportedly constructed table of test specifications than did the instructors of other disciplines of knowledge.

As clearly indicated in the literature, the variability in teachers' use of procedures for student assessment is a function of their knowledge about testing and measurement (Daniel & King, 1998; Schafer & Lissitz, 1987; Wolf, 1995). Thus, the participants' inadequate knowledge about testing and measurement could have been the single most important contributor to their failure to develop valid tests. That it is more likely for an instructor of psychology, education, and other human sciences to develop table of specifications for their tests lends weight to the argument for the importance of procedural knowledge in assessment. It is a common belief that these areas of studies are relatively more concerned about testing and measurement in comparison with other disciplines of knowledge. Therefore, the instructors of psychology, education, and other areas of human sciences are more knowledgeable about student assessment, and accordingly they are more likely to apply the fundamental procedure of creating the test blueprint.

It is also interesting to note that the participants used more of the procedures for reliable scoring than they did the procedures for developing valid achievement tests. In short, university instructors seem to be concerned more about the reliability of their tests. However, this finding is not surprising. The fact is that it is the university policy to emphasize and regulate certain examination procedures, among which the procedures require instructors to submit their tests along with the scoring plan to their respective department for vetting purposes. Such a policy, presumably, increases instructors' awareness and knowledge about testing and measurement procedures, and thus their use of reliable scoring procedures. If the argument is valid, then it is reasonable to suggest that administrative intervention is useful as a means to promote instructors' procedural knowledge and practices in student assessment.

Surprisingly, however, instructors' participation in educational development program did not produce corresponding impact on their practices in assessing students' achievement. It is equally likely for an instructor to create table of specifications, performance objectives, and scoring plan regardless of whether or not he or she had attended the program. One possible explanation for this outcome is that the educational development program was inadequate to enhance the instructor's knowledge on student assessment. More precisely, a three-hour session to discuss issues related to classroom testing and measurement is insufficient to make an impact on one's understanding about student assessment, let alone to positively change one's practices. This finding is in fact consistent with the results of earlier works on teachers' testing practices (Daniel & King, 1998; Schafer & Lissitz, 1987; Roeder, 1972). Roeder (1972), for example, claimed that teachers were still incompetent in assessing students' achievement, in spite of having received some training in classroom testing and measurement. It is imperative, therefore, for future studies to shed some light on this issue, especially with respect to the structure, the contents, and the duration of training in assessment to cater to the professional needs of teachers and university instructors.

In conclusion, confined within the limitations of the present study, the results add information to the body of knowledge pertaining to an integral component of effective teaching–student assessment as practiced by university instructors. In light of this minor contribution, more efforts are needed to help school and university teachers to master the knowledge and skills of testing and measurement.

REFERENCES

ANASTASI, A. 1988. *Psychological testing* (sixth ed.). New York: Macmillan.

CANANDY, R.L., & HOTCHKISS, P.R. 1989. It's a good score! Just a bad grade. *Phi Delta Kappan* **71:** 68-71.

COFFMAN, W.E. 1971. Essay examination. In *Educational Measurement* 2nd ed. R.L. Thorndike (Ed.), Washington, D.C. :American Council on Education.

DANIEL, L.G., & KING, D.A. 1998. Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *The Journal of Educational Research*, **91(6):** 331-344.

DIAMOND, R.M. 1998. *Designing and Assessing Courses and Curricular: A Practical Guide.* San Francisco: Jossey-Bass.

FREEMAN, R., & LEWIS, R. 1998. *Planning and Implementing Assessment.* London: Kogan Page.

GAGNE, R.M., BRIGGS, L.J., & WAGER, W. W. 1988. *Principles of Instructional Design.* New York: Holt, Rinehart, & Winston.

GOSLIN, D.A. 1967. *Teachers and Testing.* Hartford, CT: Connecticut Printers.

GRONLUND, N.E. 1985. *Measurement and Evaluation in classroom.* 5th ed. New York: Macmillan.

GULLICKSON, A.R. 1984. Teacher perspective of their instructional use of tests. *The Journal of Educational Research* **77(4):** 244-248.

GULLICKSON, A.R. 1985. Student evaluation techniques and their relationship to grade and curriculum. *The Journal for Educational Research* **79 (2):** 96-100.

HILLS, J.R. 1991. Apathy concerning grading and testing. *Phi Delta Kappan* **72:** 540-545.

International Islamic University Malaysia. *Management Services Division.* 1999. Statistics on academic staff, International Islamic University Malaysia. Unpublished data.

Joint Committee on Testing Practices. 1988. *Code of Fair Testing Practices in Education.* Washington, D.C.

KRAEMER, H.C., & THIEMANN, S. 1987. *How many subjects? Statistical Power Analysis in Research.* Newbury Park, CA: Sage.

Ministry of Education. 1999. *Diploma in Education/ Post-Graduate Pre-service Teacher Program: A guide for 1999/200 academic session.* [Buku panduan Diploma Pendidikan/Kursus Perguruan Lerpas Ijazah sesi akademik 1999/2000]. Kuala Lumpur, Malaysia. Division of Higher Learning, Ministry of Education.

OOSTERHOF, A.C. 1990. *Classroom Application of Educational Measurement.* Columbus, Ohio: Merrill.

POPHAM, W.J. 1995. *Classroom Assessment: What Teachers Need to Know.* Boston: Allyn-Bacon.

MESSICK, S. 1992. Validity of test interpretation and use. In *Encyclopedia of Educational Research.* M.C. Alkin (Ed.) 6th ed. p.1487-1495. New York: Macmillan.

NOLL, V.H. 1955. Requirements in educational measurements for prospective teachers. *School and Society* **82:** 88-90.

PALOMBA, C.A., & BANTA, T.W. 1999. *Assessment Essentials: Planning, Implementing and Improving Assessment in Higher Education.* San Francisco: Jossey-Bass.

ROEDER, H.H. 1972. Are today's teachers prepared to use tests? *Peabody Journal of Education* **49:** 239-240.

SCHAFER, W.d., & LISSITZ, R.W. 1987. Measurement training for school personnel: Recommendation and reality. *Journal of Teacher Education* **38(3):** 57-63.

SEDLEMEYER, P., & GIGERENZER, G. 1989. Do studies of statistical-power have an effect on the power studies? *Psychological Bulletin* **105:** 309-316.

SHAPIRO, B.C. 1995. The NBPTS sets standards for accomplished teaching. *Educational Leadership* **52 (2):** 55-57.

SHAEAFFER, R.L., MENDENHALL, W., & OTT, R.L. 1996. *Elementary survey sampling.* 5th. ed. Belmont, CA: Wadsworth.

SHIFFLETT, B., PHIBBS, K., & SAGE, M. 1997. Attitudes toward collegiate classroom testing. *Educational Research Quarterly* **21:** 15-26.

STIGGINS, R.J. 1991. Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice* **11(2):** 35-39.

STIGGINS, R.J., & CONKLIN, N.F. 1988. *Teacher Training in Assessment.* Portland, OR: Northwest Regional Educational Laboratory.

WISE, S.L., LUKIN, L.E., & ROOS, L.L. 1991. Teacher beliefs about training in testing and measurement. *Journal of Teacher Education* **42(1):** 37-42.

WOLF, K.P. 1995. Assessment quotes. *Reading Today* **12(4):** 4.

(Received: 15 July 1999)