International Education Studies; Vol. 8, No. 4; 2015
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Assessment for Learning Instrumentation in Higher Education

Erwin Akib[1] & Mohamad Najib Abdul Ghafar[1]

[1] Faculty of Education, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Correspondence: Erwin Akib, Faculty of Education, Universiti Teknologi Malaysia, Johor Bahru, Malaysia. Tel: 60-111-616-1796. E-mail: erwinakib@yahoo.com

Received: January 22, 2015 Accepted: February 22, 2015 Online Published: March 30, 2015

doi:10.5539/ies.v8n4p166 URL: http://dx.doi.org/10.5539/ies.v8n4p166

Abstract

This study explains assessment for learning instrumentation, especially in higher education. The population of this study was 100 lecturers of the Muhammadiyah Makassar University, Indonesia. A total of 50 items from six construct were analyzed and used to determine the reliability and validity of the questionnaire. The result shows that the person reliability of the instrument of 100 people was 0.91. It showed that the person reliability is excellent, as well item reliability showed a valued 0.96, which can be categorized excellent. It suggests that the goals of assessment should be to encourage that universities are administered in a way that provides the most appropriate practice in developing teaching and learning process.

Keywords: assessment for learning, instrumentation, higher education

1. Introduction

In an effort to run the article 31 of Law 1945, the Indonesian government from time to time continue to make the development of education through the development of a national education system. The national education system is the overall educational components are interlinked in an integrated manner to achieve national education goals. In the development of national education system can never be separated from the color of the social, political, economic and culture that surrounds them. From the perspective of the national education system, we recognize the national education system version of the old order, the new order, and the order of reform.

The various problems encountered in improving the quality of education in Indonesia, began from primary education to higher education. This is in accordance with what is expressed by Ramly (2005) mentions some of the critical issues of education in Indonesia, among others: the strike of teachers, Higher Education Accreditation System is commercial, the evaluation system is not accommodating, the influx of foreign investment in education, providing education for local authority that the irregularities, the ability of teachers weak in mastering teaching materials, educational institutions and become a contributor of educated unemployment, sectoral egoism meterialismedan scientists, education becomes cheap business arena, and the occurrence of educational teaching materials not only control the behavior and moral development and the absence of taxes for education.

Teaching and learning process does not only talk about the process, but it also talks about the results. Hence, to know the outcome of that process, teachers or lecturers should use the test as a tool in measuring the students' ability or performance, and decided, whether the students can pass or not. In the process of teaching and learning, lecturers not only focus on the teaching process, but also on how they measure their students or apprentices outcomes. Reynolds et al. (2010) stated that the assessment is a systematic process to gather information that can be used to draw conclusions about objects or processes. Ghafar (2011) explained that the assessment is a systematic procedure that involves the collection, analysis and translation of evidence that the student has reached as far as teaching purposes occurs. A number of authors have reported a negative impact of assessment on learning and teaching (Frederiksen, 1984; Ridgway & Schoenfeld, 1994; Dochy & McDowell, 1997). This case demonstrates that assessment has significant impact on teaching and learning.

Ghafar (1999, 2011) explains that the reliability refers to the consistency of test results. If a person has a certain skill level, she or he is able to demonstrate the same level when retested, the skill level is reliable. Reliability can be determined by the test-retest, split half, equivalent for parallel, Kuder Richardson, inter-examiner, and inter-observer methods (Ghafar, 1999, 2011; Creswell, 2012; Fraenkel & Wallen, 2009). Reliability is an important issue in the use of any instrument if the instrument had been used in other research or if the instrument

is built for the purpose of research. Validity is most important when preparing or selecting an instrument. Researchers intend to obtain information using an instrument. Validity include types of measures and procedures of measurement, including formal tests, observation techniques, interview protocols, questionnaires, self-report affective measures, projective devices, and so on (Ghafar, 1999, 2011; Goodwin, 2002). The term validity includes two aspects, what is to be measured and how consistently it is measured (Ebel & Frisbie, 1991).

Historically, the term "assessment for learning" begins with the term formative assessment that includes an assessment for learning has been observed by Black and Wiliam (2006) and Newton (2007) from writing Scriven (1967) first distinguishes the difference between formative and summative assessment purposes, the work of Bloom, Hasting and Madaus (1971) and the work of Sadler (1989), which highlights the importance of formative set criteria to inform students about learning.

Important assessment for learning research works for teachers and students has begun in the UK (Black, Swann, & Wiliam, 2006; Ecclestone, 2002; Gardner et al., 2008; Gipps, 2002; Hayward, 2007; Marshall & Drummond, 2006; Stobart, 2009) the USA (Brokhart, 2001; Popham, 2008; Stiggins, 2002; Tierney & Charland, 2007) Hong Kong (Carless, 2007), New Zealand (Cowie, 2005b; Hattie & Tumpeley, 2007) and in other places around the world.

The focus of assessment for learning is increasing students' achievement (Reeves, 2001) and the students learn rather than teaching (Harris, 2007). Assessment for learning also includes the feedback designed to provide immediate, relevant and useful information to students and the formative feedback aims to provide information communicated to the students to support the modification of thought or behavior to improving learning (Shute, 2008).

Assessment for learning relate to practices, such as sharing criteria with students, developing a classroom talk and asking questions, providing appropriate feedback, and allowing peer and self-assessment (Black and Wiliam 1998a) all requiring the active involvement of students. Learning is seen as a process rather than a product (Sadler, 2007). Teachers need to provide opportunities for students to learn to understand and to engage in thoughtful discussion. Students are not passive recipients of knowledge. They have become their own learning controller for self-assessment and peer assessment. Carless (2005) showed the two cases for the implementation of assessment for learning in Hong Kong. One of the cases that show how an English teacher in primary schools share the assessment criteria with the students and the students grab a part in assessing their peers using a checklist.

Additional cases reported how an English teacher incorporated evaluating peer in the classroom in order to increase student grammar. According to this study (James & Pedder, 2006; Keppell & Carless, 2006; Marshall & Drummond, 2006; Munns & Woodward, 2006) uses the implementation of assessment for learning as a pedagogical training is far additional complex. Bernstein (cited in Munns & Woodward, 2006) provides a lens that displays the subject of interpretation influential educator and student beliefs, personality and manipulation to help understand the complexity. Moreover, in the context of society and a very important strategy in assessment for learning and never linear and closed as a series of relate above. This is while training can inform theory assessment for learning. Reality that teachers and students debate can help researchers explain and understand the dynamics of the relationship assessment. Furthermore, Stiggins (2004, 2006) stated that assessment for learning argues that students learn best when they know what is expected and required for success, and they understand how to close the gap between their own work and the standard for success. The strategy in providing students with this knowledge about what is expected can be found in the use of scoring guide. Accessible instructional scoring guides or rubrics can provide students with important information that can lead students to become self-regulated learner (Saddler and Andrade, 2004).

2. Research Objective

A research was carried out with the objective to investigate the assessment for learning instrumentation in higher education.

3. Methodology

The research design utilized was the descriptive survey design, involving only a one-time response to the questionnaire. Fraenkel and Wallen (2009) explained that survey research is intended to obtain data to determine specific characteristics of a group. The Rasch model analysis is used as a tool to know the reliability of the instruments. The items used are the Likert scale type totaling 50 items. The questions were formulated based on six constructs for Assessment for Learning.

3.1 Design of Instrument

The constructs and construct indicators or items of the questionnaire were divided into six constructs which are Sharing Learning Objectives (SLO), Helping Pupils (HP), Peer and Self-Assessment (PSA), Providing Feedback (PF), Promoting Confidence (PC), and Involving in Reviewing and Reflecting (IRR).

3.2 Population

The populations of this research were all of the lecturers of University of Muhammadiyah, which has education faculty in Indonesia, and the sampling technique used purposive sampling, therefore the number of samples was 100 lecturers at the faculty of education of University Muhammadiyah of Makassar, South Sulawesi, Indonesia.

3.3 Validation of Instrument

The instrument validation involved four steps: (i) metadata analysis, (ii) expert validation, (iii) pilot test, and (iv) data analysis using the Rasch Measurement Model with Winstep software. After completing the metadata analysis, the instrument was validated for constructing and content validity of expert in Measurement and Evaluation, Faculty of Education UTM and for face validity for by expert in Language education of the Makassar Muhammadiyah University for face validity. After correcting the instrument as suggested, the pilot study was conducted. Finally, the data were analyzed measure the validity and reliability using the Rasch Measurement Model.

3.4 Data Analysis

A total of 50 items from six construct were analyzed and used to determine the reliability and validity of the questionnaire. Statements were coded as numerical responses with Likert Scale rather than as words or phrases. All data were verified by hand checking, coded numerically, and entered onto the SPSS version 20. The analysis using RASCH Model with Winstep software for validation process was then carried out.

4. Findings

The first step is to analyze the questionnaire whether some items needed to be deleted or modified. The reliability and validity of the questionnaire were measured using person reliability, item reliability, item dimensionality, and difficulty level of scales. In the person misfit table, the columns that needed to be observe were Pt-Measure Corr., outfit MNSQ and Z-STD, and infit MNSQ) and ZSTD (Azrilah, 1996). If the outfit MNSQ and Z-Std value is large, but the infit MNSQ and ZSTD value is within the range, the misfit is still acceptable because of the sloppy response (Azrilah, 1996).

4.1 Person Reliability

The person reliability of the instrument of 100 people was 0.91. It showed that the person reliability is excellent (Fisher, 2007). After deleting 26 responds, the Rasch analysis has conducted for the other 74 responds. Person reliability, increased from 0.91 to 0.94. It indicated that the reliability of the instrument was still within the excellent category (Fisher, 2007), as shown in the Table 1.

Table 1. Person reliability for 100 respondents

	Total Score			Model Error	Infit		Outfit	
		Count	Measure		MNSQ	ZSTD	MNSQ	ZSTD
MEAN	206.3	50.0	2.46	.25	1.04	.0	.99	-.1
S.D.	13.9	.0	.87	.02	.31	1.5	.30	1.5
MAX.	231.0	50.0	4.16	.30	2.08	4.1	1.97	3.8
MIN.	173.0	50.0	.65	.21	.47	-3.1	.44	-3.5
REAL RMSE .27	TRUE SD .83	SEPARATION 3.14	PERSON RELIABILTY .91					
REAL RMSE .25	TRUE SD .84	SEPARATION 3.33	PERSON RELIABILTY .92					
S.E. OF PERSON MEAN = .09								

Table 2. The person reliability, after deleting 26 respondents

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	205.9	50.0	3.03	.27	1.01	.0	.99	-.1
S.D.	14.9	.0	1.09	.02	.26	1.3	.25	1.3
MAX.	231.0	50.0	5.00	.32	1.48	2.1	1.40	1.7
MIN.	173.0	50.0	.85	.24	.45	-3.3	.41	-3.6
REAL RMSE .29	TRUE SD 1.05	SEPARATION 3.67	PERSON RELIABILTY . 93					
REAL RMSE .27	TRUE SD 1.05	SEPARATION 3.87	PERSON RELIABILTY .94					
S.E. OF PERSON MEAN = .13								

DELETED: 26 PERSON

PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00

CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .94

4.2 Item Reliability

Item reliability showed a valued 0.96, which can be categorized excellent (Fisher, 2007). The misfits' pattern to be considered were focused on the three (3) columns, that are 0.4 <Point Measure Correlation (PtMea Corr) value <0.85, 0.5 <outfit Mean Square (MNSQ) value <1.5, and -2 <outfit Z-Standard (ZSTD) value <+2 (Azrilah, 1996).

Table 3. Item reliability for 74 respondents

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	304.7	74.0	.00	.22	.99	-.3	.99	-.4
S.D.	25.5	.0	1.21	.02	.40	2.3	.41	2.3
MAX.	342.0	74.0	2.45	.25	1.90	4.2	1.95	4.3
MIN.	249.0	74.0	-1.98	.19	.35	-5.3	.34	-5.4
REAL RMSE .24	TRUE SD 1.19	SEPARATION 4.95	ITEM RELIABILTY . 96					
REAL RMSE .22	TRUE SD 1.19	SEPARATION 5.31	ITEM RELIABILTY .97					
S.E. OF PERSON MEAN = .17								

4.3 Item Validity

Table 4 indicates the scale of 40 persons. There are five (5) scales. They were Strongly Agree (SA), Agree (A), Uncertain (U), Disagree (D), and Strongly Disagree (SD). In the Rasch measurement model, the differences between each ranking are taken into account. The difference must be in the range of 1.5<s<5.0 (Azrilah, 1996).

Table 4. Scale calibration of 74 persons

SUMMARY OF CATEGORY STRUCTURE. Model="R"

Category Label	Score	Observed Count	%	Observed Average	Sample Expect	Infit MNSQ	Outfit MNSQ	Structure Calibratin	Category Measure	
1	1	3	0	.88	-.74	1.87	2.25	NONE	-4.68	1
2	2	67	2	.24	.01	1.16	1.29	-3.49	-2.49	2
3	3	486	13	1.17	1.24	.94	.93	-1.40	-37	3
4	4	2080	56	2.86	2.86	.96	.91	.59	2.47	4
5	5	1064	29	4.39	4.37	1.04	1.00	4.30	5.42	5

OBSERVED AVERAGE is mean of measures in the category. It is not a parameter estimate.

In Rasch Measurement Model, the probability of responses, whether the scales are equally distributed can be measured or using the scale calibration. Calibration scale is designed to identify the level of difficulty of the questionnaire on the grading scale. It is mandatory to have respondents' information in terms of their ability in distinguishing the scale rating. It was found that the scale differences scales were more than 1.5 and less than 5 except in scale 2 (Disagree) and 5 (Strongly Agree). This indicated that the respondents found difficulty to distinguish the scale 2 (Disagree) and scale 5 (Strongly Agree).

5. Conclusion

This study showed that the person reliability was categorized as fair, but the item reliability was as Excellent, and the respondents found difficulty to distinguish the scale 2 (Disagree) and scale 5 (Strongly Agree). This study shows the importance of considering symmetry measures due to the gap between person reliability, item reliability, and difficulty level of scales.

References

Azrilah, A. A. (1996). *Rasch model fundamentals: Scale constructs and measurement structure*. Integrated Advance Planning Sdn. Bhd.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability*. Chicago: University of Chicago Press.

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81-100). London: Sage.

Bloom, B. S., Hasting, J. T., & Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. McGraw-Hill Book Co, New York.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education, 8*(2), 153-169. http://dx.doi.org/10.1080/09695940123775

Carless, D. (2007a). Learning-oriented assessment: Conceptual basis and practical implications. *Innovations in Education and Teaching International, 44*(1), 57-66. http://dx.doi.org/10.1080/14703290601081332

Carless, D. (2007b). Conceptualizing pre-emptive formative assessment. *Assessment in Education, 14*(2), 171-184. http://dx.doi.org/10.1080/09695940701478412

Cowie, B. (2005b). Pupil commentary on assessment for learning. *Curriculum Journal, 16*(2), 137-151. http://dx.doi.org/10.1080/09585170500135921

Creswell, J. W. (2012). *Educational Research (Planning, Conducting, and Evaluating Quantitative and Qualitative Research)* (4th ed.). Boston, USA: Pearson.

Dochy, F. J. R. C., & McDowell, L. (1997). Introduction assessment as a tool for learning. *Studies in Educational Evaluation, 23*(4), 279-298. http://dx.doi.org/10.1016/S0191-491X(97)86211-6

Dochy, F., Segers, M. S. R., & Sluijsmans, D. (1999). The use of self-, peer and coassessment in higher education: A literature review. *Studies in Higher Education, 24*(3), 331-350. http://dx.doi.org/10.1080/03075079912331379935

Ebel, R. L., & Frisbie, D. A. (1990). *Essentials of Educational Measurement* (5th ed.). Englewood Cliffs, N.J.: Prentice-Hall.

Ecclestone, K. (2002). *Learning autonomy in post-compulsory education: The politics and practice of formative assessment*. London: Routledge.

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Fraenkel, J. R., & Wallen, N. E. (2009). How to Design and Evaluate Research in Education. *Qualitative Research* (7th ed.). McGraw-Hill Higher Education.

Frederiksen, J. R., & Collins, A. (1989). A systematic approach to educational testing. *Educational researcher, 18*(9), 27-32. http://dx.doi.org/10.3102/0013189X018009027

Gardner, J. et al. (2008). *Changing assessment practice: Process, principles and standards*. London: Assessment Reform Group.

Ghafar, M. N. A. (1999). *Penyelidikan Pendidikan*. Skudai: Penerbit Universiti Teknologi Pendidikan Malaysia.

Ghafar, M. N. A. (2003). *Reka Bentuk Tinjauan Soal Selidik Pendidikan*. Skudai: Penerbit UTM.

Ghafar, M. N. A. (2011). *Pembinaan & Analisis Ujian Bilik Darjah*. Edisi Kedua. Skudai: Penerbit UTM Press.

Gipps, C. (2002). Sociocultural perspectives on assessment. In G. Wells, & G. Claxton (Eds.), *Learning for life in the 21st century* (pp. 73-83). Oxford: Blackwell publishers. http://dx.doi.org/10.1002/9780470753545.ch6

Goodwin, L. D. (2002). Changing conceptions of measurement validity: An update on the new Standards. *Journal of Nursing Education, 41*(3), 100-106.

Harris, L. (2007). Employing formative assessment in the classroom. *Improving Schools, 10*(3), 249-260. http://dx.doi.org/10.1177/1365480207082558

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. http://dx.doi.org/10.3102/003465430298487

Hayward, L. (2007). Curriculum, pedagogies and assessment in Scotland: The quest for social justice. 'Ah kent yir faither'. *Assessment in Education: Principles, Policy & Practice, 14*(2), 251-268.

Kimberlin, C., & Winterstein, A. G. (2008). Fundamentals Validity and reliability of measurement instruments used in research. *Am J Health-Syst Pharm* (Vol. 65). http://dx.doi.org/10.2146/ajhp070364

Linacre, J. M. (2014). *Reliability and separation of measures*. Winsteps Help.

Marshall, B., & Drummond, M. (2006). How teachers engage with Assessment for Learning: Lessons from the classroom. *Research Papers in Education, 21*(2), 133-149. http://dx.doi.org/10.1080/02671520600615638

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education, 14*(2), 149-170. http://dx.doi.org/10.1080/09695940701478321

Perkins, K., Wright, B. D., & Dorsey. (2000). Multiple regression via measurement [diagnosing gout]. *Rasch Measurement Transactions, 14*(1), 729-730

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Ramly, N. (2005). *Membangun Pendidikan yang Memberdayakan dan Mencerahkan*. Jakarta: Grafindo.

Reeves, D. B. (2001). Standards make a difference: The influence of standards in classroom assessment. *NASSP Bulletin, 85*(5), 5-12. http://dx.doi.org/10.1177/019263650108562102

Ridgway, J., & Schoenfeld, A. H. (1994). *Balanced Assessment: Designing Assessment Schemes to Promote Desirable Change in Mathematics Education*. Keynote paper for the EARLI Email Conference on Assessment.

Sadler, D. R. (1989) Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144. http://dx.doi.org/10.1007/BF00117714

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy, and Practice, 5*, 77-84. http://dx.doi.org/10.1080/0969595980050104

Sadler, D. R. (2007). Beyond feedback, developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*, 535-550. http://dx.doi.org/10.1080/02602930903541015

Scriven, M. (1967). *The methodology of evaluation* (Washington, DC, American Educational Research Association).

Scriven, M. (2002). Evaluation ideologies. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation: Viewpoints on educational and human services evaluation models* (2nd ed., pp. 249-278). Boston, MA: Kluwer Academic Publishers.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(2), 153-189. http://dx.doi.org/10.3102/0034654307313795

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758-765. http://dx.doi.org/10.1177/003172170208301010

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, *51*(2), 161-179. http://dx.doi.org/10.1080/00131880902891305

Tierney, R. D., & Charland, J. (2007, April). *Stocks and prospects: Research on formative assessment in secondary classrooms*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED496236)

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Zessoules, R., & Gardner, H. (1991). Authentic assessment: Beyond the buzzword and into the classroom. In V. Perrone (Ed.), *Expanding student assessment* (pp. 47-71). Alexandria, VA: Association for Supervision and Curriculum Development.