Education Journal
2015; 4(2): 64-68
Published online March 4, 2015 (http://www.sciencepublishinggroup.com/j/edu)
doi: 10.11648/j.edu.20150402.13
ISSN: 2327-2600 (Print); ISSN: 2327-2619 (Online)

The Validity and Reliability of Assessment for Learning (AfL)

Erwin Akib, Mohamed Najib Abdul Ghafar

Measurement and Evaluation, Faculty of Education, Universiti Teknologi Malaysia, Johor, Malaysia

Email address:
erwinakib@yahoo.com (E. Akib)

To cite this article:
Erwin Akib, Mohamed Najib Abdul Ghafar. The Validity and Reliability of Assessment for Learning (AfL). *Education Journal.* Vol. 4, No. 2, 2015, pp. 64-68. doi: 10.11648/j.edu.20150402.13

Abstract: Assessment for learning is a new perspective on the assessment system in education. The traditional practice is for evaluating outcomes is an Assessment of Learning. However, new perspective proposes that assessment should be included in the process of learning, that is Assessment for Learning. This main objective of this study is to investigate the validity and reliability of Assessment for Learning. This study used the quantitative survey design, carried out in Indonesia using the proportional stratified random sampling method involving 100 lecturers. It was conducted at University Muhammadiyah of Makassar, South Sulawesi, Indonesia. The data were analyzed using: t-test, anova, and chi-square. The instrument validity and reliability were determined using Rash model analysis. The finding shows that the validity and reliabity of each construct of Assessment for Learning has a high level.

Keywords: Assessment for Learning, Reliability, Validity

1. Introduction

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 Nadjamuddin 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, 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. Mohd. Najib (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 and Schoenfeld, 1994; Dochy and McDowell, 1997). This case demonstrates that assessment has significant impact on teaching and learning.

Najib (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 (Najib, 1999; Najib, 2011; Creswell, 2012; Fraenkel and 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 (Najib, 1999; Najib, 2011; Goodwin, 2002). The term validity includes two aspects, what is to be measured and how consistently it is measured (Ebel and 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 & 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. Many educators have promoted the use of alternative assessment, assessment for learning that could reveal students' learning processes better than traditional assessments that only focus on students' learning outcomes (Zessoules & Gardner, 1991; Wiggins, 1998). They believe that the use of assessment for learning in classroom instruction can empower students as learners and thus improve students' performance (Sadler, 1998; Black et al., 2004).

Important assessment for learning research works for teachers and students has begun in the U.K (Black, Swann, & Wiliam, 2006; Ecclestone, 2002; Gardner et al., 2008); Gipps, 2002; Hayward, 2007; Marshall & Drummond, 2006; Stobart, 2009) the U.S.A (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.

However, assessment for learning has yet to show the level of validity and reliability are adequate, especially in higher education in Indonesia and more specifically related to the understanding of the lecturer. Therefore, this study will describe details about the validity and reliability of assessment for learning.

2. Objective of the Study

This study was aimed to investigate the validity and reliability of Assessment for Learning at University of Muhammadiyah Makassar, South Sulawesi Indonesia.

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. This study involved 100 lecturers at the University of Muhammadiyah Makassar, South Sulawesi, Indonesia.

3.1. The Design of Instrument

The constructs and construct indicators or items of the questionnaire were divided into six constructs which are Sharing Learning Objectives (SLO) that consisted of 12 items, Helping Pupils (HP) consisted of 7 items, Peer and Self – Assessment (PSA) consisted of 9 items, Providing Feedback (PF) consisted of 8 items, Promoting Confidence (PC) consisted of 6 items, and Involving in Reviewing and Reflecting (IRR) consisted of 8 items. In constructing the instrument there are several issues to consider. First is the identification of variables in the research. The next step is to write an operational definition for each of the attributes that has been assigned by certain keywords, so that the attributes appear more meaningful and reflective of what the attributes is about (Azrillah, 1996). 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.

3.2. 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 observed 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 respondent (Azrilah, 1996).

4.1. Person Reliability

One way to think of reliability is that, other things being equal, a person should get the same score on a questionnaire if they complete it at two different points in time (test-retest reliability. Another way to look at the reliability is to say that two people who are the same in terms of the construct being measured, should get the same score. In statistical terms, the usual way to look at reliability is based on the idea that individual items (or sets of items) should produce results consistent with the overall questionnaire. The reliability estimates are used to evaluate (1) the stability of measures administered at different times to the same individuals or using the same standard (test–retest reliability) or (2) the equivalence of sets of items from the same test (internal consistency) or of different observers scoring a behavior or event using the same instrument (interrater reliability). Reliability coefficients range from 0.00 to 1.00, with higher coefficients indicating higher levels of 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 and 2 below.

Table 1. *Person Reliability for 100 Respondents*

	TOTAL		MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	206.3	50.0	2.46	.25	1.01	.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 RELIABILITY	.91	
MODEL RMSE	.25	TRUE SD	.84	SEPARATION	3.33	PERSON RELIABILITY	.92	
S.E. OF PERSON MEAN = .09								

PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .91

Table 2. *The person reliability, after deleting 26 respondents*

	TOTAL		MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	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 RELIABILITY	.93	
MODEL RMSE	.27	TRUE SD	1.05	SEPARATION	3.87	PERSON RELIABILITY	.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		MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	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	1.98	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	PERSON RELIABILITY	.96	
MODEL RMSE	.22	TRUE SD	1.19	SEPARATION	5.31	PERSON RELIABILITY	.97	
S.E. OF PERSON MEAN = .17								

4.3. Item Validity

Validity is often defined as the extent to which an instrument measures what it purports to measure. Validity requires that an instrument is reliable, but an instrument can be reliable without being valid. Table 4 indicates the scale of 40 persons. There are five (5) scales. They are 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*

CATEGORY		OBSERVED		OBSVD	SAMPLE	INFIT	OUTFIT	STRUCTURE	CATEGORY	
LABEL	SCORE	COUNT	%	AVRGE	EXPECT	MNSQ	MNSQ	CALIBRATN	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

SUMMARY OF CATEGORY STRUCTURE. Model="R"
OBSERVED AVERAGE is mean of measures in 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. Rasch analysis can help to determine the validity of the scale was used to make the determination of zero and then making the calibration scale is used. Rasch analysis to determine the validity of the probability of response is spread evenly between the scales of the fixed (Norlide, 2007; Azrilah Aziz, 2010 and Perkins et al., 2002). 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

[1] N. Ramly. Membangun Pendidikan yang Memberdayakan dan Mencerahkan. Jakarta: Grafind, 2005..

[2] M. N. Ghafar, Pembinaan & Analisis Ujian Bilik Darjah. Edisi Kedua. Skudai: Penerbit UTM Press, 2001.

[3] J.R. Frederiksen, and A. Collins, A systematic approach to educational testing. Educational researcher, 1989. 18 (9), pp. 27-32.

[4] J. Ridgway and A. H. Schoenfeld, Balanced Assessment: Designing Assessment Schemes to Promote Desirable Change in Mathematics Education. Keynote paper for the EARLI Email Conference on Assessment, 1994.

[5] F.J.R.C. Dochy and L. McDowell, Introduction assessment as a tool for learning. Studies in Educational Evaluation, 1997.. Vol. 23, No. 4, pp. 279 – 298.

[6] M.N. Ghafar, Penyelidikan Pendidikan. Skudai: Penerbit Universiti Teknologi Pendidikan Malaysia, 1999.

[7] J.W. Creswell, Educational Research (Planning, Conducting, and Evaluating Quantitative and Qualitative Research). Fourth Edition. Boston, USA: Pearson, 2012.

[8] F.J. Fraenkel and N.E. Wallen, How to Design and Evaluate Research in Education. Qualitative Research (7th ed.). McGraw-Hill Higher Education, 2009.

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

[10] R.l. Ebel and D.A. Frisbie, Essentials of Educational Measurement. (5th Ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1990.

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

[12] P. E. Newton, Clarifying the purposes of educational assessment, Assessment in Education, July 2007, Vol. 14, No. 2 pp. 149–170.

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

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

[15] D.R. Sadler, Formative assessment and the design of instructional systems, Instructional Science, 1989, pp. 119–144.

[16] R. Zessoules and H. Gardner, 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, 1991.

[17] G. Wiggins, Educative assessment : Designing assessments to inform and improve student performance. San Francisco, CA: Jossey-Bass, 1998.

[18] D.R. Sadler, Formative assessment: Revisiting the territory. Assessment in Education: Principles, Policy, and Practice,1998, pp. 77–84.

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

[20] K. Ecclestone, Learning autonomy in post-compulsory education: The politics and practice of formative assessment. London: Routledge, 2002.

[21] J. Gardner, et al, Changing assessment practice: process, principles and standards. [London]: Assessment Reform Group, 2008.

[22] C. Gipps, Sociocultural perspectives on assessment. In Learning for life in the 21st century, eds. G. Wells and G. Claxton (Eds.), 73 – 83. Oxford: Blackwell publishers, 2002.

[23] L. Hayward, Curriculum, pedagogies and assessment in Scotland: The quest for social justice. 'Ah kent yir faither'. Assessment in Education: Principles, Policy & Practice, 2007, pp. 251–68

[24] B. Marshall, and M. Drummond, How teachers engage with Assessment for Learning: lessons from the classroom. Research Papers in Education 21, 2006, no. 2: pp. 133 - 149

[25] G. Stobart, Determining validity in national curriculum assessments.Educational Research, 2009, 51(2), pp. 161-179.

[26] S.M. Brookhart, Successful students' formative and summative uses of assessment information. Assessment in Education, 2001, 8 (2): pp. 153-169.

[27] W.J. Popham, Transformative assessment. Alexandria, VA: Association for Supervision and Curriculum Development, 2008.

[28] R.J. Stiggins, Assessment crisis: The absence of assessment for learning. Phi Delta Kappan, 2002, Vol. 83, No. 10, pp. 758–765.

[29] R.D. Tierney and J.Charland, 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), 2007.

[30] D. Carless, Learning-oriented assessment: Conceptual basis and practical implications. Innovations in Education and Teaching International, 2007, 44(1), pp.57-66.

[31] B. Cowie, Pupil commentary on assessment for learning. Curriculum Journal, 2005b, 16(2), pp. 137-151.

[32] J. Hattie and H. Timperley, The power of feedback. Review of Educational Research, 2007, 77(1), pp. 81-112.

[33] D.B. Reeves, Standards make a difference: The influence of standards in classroom assessment. NASSP Bulletin, 2001, 85(5), pp. 5- 12.

[34] L. Harris, Employing formative assessment in the classroom. Improving Schools, 2007, 10(3), pp. 249-260.

[35] V.J. Shute, Focus on formative feedback. Review of Educational Research, 2008, 78(2), pp. 153-189.

[36] A.A. Azrilah, Rasch model fundamentals: scale constructs and measurement structure. Integrated Advance Planning Sdn. Bhd, 1996.

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

[38] Norlide, Using The Rasch Measurement Model For Standard Setting Of The English Language Placement Test At The IIUM. Tesis Ijazah Doktor Falsafah. Universiti Islam Antarabangsa, 2007.

[39] K. Perkins, B.D. Wright and Dorsey, Multiple regression via measurement [diagnosing gout]. Rasch Measurement Transactions, 2002, 14(1), 729-30