Investigating the Quality of a Popular Classroom Assessment Instrument in Indonesia

Taufiq Effendi, Ichwan Suyudi
Universitas Gunadarma
taufiq.effendi@gmail.com

Abstract - **This paper reports a preliminary study on the quality of a multiple-choice test as a popular classroom assessment in Indonesia. An item response analysis was conducted to scrutinize four major kinds of quality: reliability, construct validity, item discrimination ability, and distractor plausibility. The data were a form of test developed by a local teacher with strong credentials and administered to roughly eighty students, the test takers' responses to each test item and the data processed by ConQuest, a program developed by Australian Council for Educational Research. The study uncovered that the test requires some reconstruction to establish adequate construct validity. At least 13.5% of all items failed to discriminate students on their ability and roughly half of all items had deficient distractors, not to mention some which distracted no students. Additionally, the test was also found to have failed to meet the acceptable cut-off score of reliability. Finally, this preliminary research highlights two important implications: first, the importance of classroom teachers to receive a proper training on assessment to be able to accurately monitor and reveal the learners' genuine learning progress, and second, the importance of conducting wider scale research to investigate teachers' capacity in developing their classroom assessment instruments.**

Keywords: *classroom assessment, multiple-choice test, validity, reliability, item response analysis*

1. INTRODUCTION

In the context of English as a Foreign Language (EFL) or a Second Language (ESL), tests in a multiple-choice (MC) format have been notorious for some of their critical weaknesses. Hughes (2012, p. 76) argues that MC tests are highly speculative; they facilitate guessing and cheating. Learners can simply select an option that they think is the most distinct, not the correct one based on their ability, which makes the interpretation of the test results problematic (Weir, 1990, p. 14). Such inaccurate representation of the level of the ability of the learners certainly leads to a deceiving picture of where the learners are in the continuum of the learning. This certainly staggers teachers to be able to properly guide and help learners to arrive at the targeted learning destination.

On the other hand, MC tests are argued to be objective in a way they do not involve marker's judgement (Miller, Linn & Gronlund, 2009, p. 155). However, the nature of the test item, the distracters and the correct answer are often based on the subjectivity of the test developer (Weir, 1990, p. 43). This is why it is common to find disagreements between the test takers and the test developers about which is the correct answer. This would be even more obvious if the test developers possess low assessment expertise and insufficient English language proficiency which is often the case (Yulia, 2014).

Although multiple-choice tests are somewhat subjective, some still believe that MC tests still hold significant importance due to their wide applicability that ranges from knowledge comprehension to its application and their cost-effective quality (Wu & Adams, 2007). This resonates the condition in Indonesian context. Multiple-choice tests are relatively popular not only as a classroom assessment instrument but also as a high stake examination (Yulia, 2014; Zulfikar, 2009). One reason for this popularity is that in Indonesia, teachers generally teach large classes for at least 24 hours a week, not to mention extra teaching hours many teachers do in other institutions in order to earn more to make a living. Therefore, to save time, many teachers favour multiple-choice tests.

Having said that, it is clear that the decision to select which assessment techniques or instruments can result in the credibility and accuracy of the information about learners' progress and achievements. For this reason, this study is conducted to investigate the quality of a multiple-choice test as one form of English classroom assessments. Clearly, this preliminary study seeks to find out how good a multiple-choice test developed by a teacher is in revealing learners' genuine ability. To guide the research, four questions are derived from the main inquiry as follows: (1) To what extent is the test result reliable? (2) To what extent does each individual test item establish its construct validity? (3) To what extent can each individual item discriminate test takers based on their ability? (4) To what degree can each distracter divert test takers?

2. METHODS

One way of investigating the quality of a multiple-choice test is by conducting an item response analysis to analyse learners' responses to the test items. In other words, it is a way to know the quality of each test item in revealing a picture of the test takers' ability as reflected through their responses to each of the items. Such analysis is based on an Item Response Theory (IRT), a subclass of Psychometric, which is aimed to discover the relationship between test takers' answers to individual test items and the position of the test takers' performance on relevant continuum (Reckase, 2009). Brown (2004) highlights that analysing item responses can help test developers to find out the weaknesses of the test and therefore have the foundation to revise the test to generate more reliable information of the test takers' ability.

2.1 Data collection and analysis

This study looked at a classroom assessment instrument in the form of a multiple-choice test on conditional clauses. The test consists of three parts that assess students' mastery of the three types of conditionals. The test was developed by a local teacher with strong credentials in the field of English Education. The test was administered to eighty school students.

Then, the study looked at the responses of the test takers to each individual item. Here, the responses given by all test takers were typed in excel document. Afterward, these responses were put into ConQuest, a program developed by the Australian Council for Educational Research.

3. FINDINGS AND DISCUSSION

3.1 Reliability

The analysis of the test results generated by ConQuest showed that the coefficient alpha or the reliability index is 0.48. The value of 0.48 is below the acceptable cut-off value of 0.70 (Schmitt, 1996 p. 351) for a cognitive test.

Reliability is generally understood as the consistency of the results of a test (Bachman & Palmer, 1996; Brown, 2004; Hughes, 2012; Weir, 2005). That means, if the value is below 0.70, it is argued that the results of the test would likely change if the same test is given to the same group of students in a different time.

To some scholars, reliability is not different from concurrent validation in which a result of an item response is measured against the result of other item response that measures the same construct (Messick, 1989). Based on this consideration, Weir (2005) prefers to take reliability as a form of validity which he calls "scoring validity". To these scholars, reliability or scoring validity significantly contributes to the overall validity or the interpretation and use of a test result. That means, it is unlikely to claim that a test produces a credible result on which interpretation and decisions are made if the reliability or other aspects of its quality is problematic. In other words, to claim that a test is valid, it requires a composite body of credible evidences. However, although scoring validity seems more appropriate, to conform to the more common term and to avoid confusion, the term "reliability" is used throughout this paper.

The reliability of the results of the test could have been influenced by both the quantity and the quality of the test items. One way to increase the reliability value of the test results is by adding more items (Weir, 1990, pp. 54-55). This suggests that the low degree of reliability could have been caused by the small number of test items. Another way to capture a more reliable evidence of the students' ability is by deleting the items that do not quite discriminate between the high and low performing students (Hughes, 2012, p. 44). That means, the low value of the coefficient alpha could have been the result of a poor quality of individual items. This suggests that test analysis literacy is vital for teachers to have in order to be able to figure out what makes the reliability of the results of a test problematic.

3.2 Construct Validity

Construct validity is an essential element of a test quality. It is generally understood as the degree of meaningfulness and adequacy of the interpretation and use of the results of a test (Bachman & Palmer, 1996; Miller, Linn & Gronlund, 2009). As the definition suggests, it provides the evidence of whether or not the test or the individual item measures the same construct which generates a representation of the test takers' ability (Bachman & Palmer, 1996; Fulcher & Davidson, 2007; Weir, 2005). Some scholars take construct validity as the umbrella for all other forms of validity which includes content-related and criterion-related validity (Messick, 1989; Miller, Linn & Gronlund, 2009). Likewise, Bachman and Palmer (1996) consider construct validity to encompass authenticity and interactiveness which according to Weir (2005) are similar to content validity or in Weir's term "context validity”, which is synonymous with content-related validity as a more common term.

In the analysis document, the degree of the construct validity of the test is labelled as "weighted MNSQ". Weighted MNSQ does not only indicate the degree of the construct validity, but it also indicates how much the item fits the item response model (Wu & Adams, 2007, p. 66). It is expected that the value is closed to 1.00 and falls within the Confidence Interval. This information is given in the following table.

Note: I is item. DI is discrimination index. D1 is distractor 1. D2 is distractor 2. D3 is distractor 3.

I	MNSQ (CI)	DI	Delta	Correct opt	Pt-bis	D1	Pt-bis	D2	Pt-bis	D3	Pt-bis
1	0.96 (0.87, 1.13)	0.49	-0.78	42 (52.50)	0.49	A: 8 (10.00)	-0.24	B: 3 (3.75)	0.05	C: 27 (33.75)	-0.38
2	1.04 (0.69, 1.31)	0.26	-2.08	63 (78.75)	0.26	A: 9 (11.25)	-0.19	C: 4 (5.00)	-0.03	D: 4 (5.00)	-0.19
3	1.04 (0.88, 1.12)	0.29	-0.62	39 (48.75)	0.29	B: 13 (16.25)	-0.05	C: 13 (16.25)	-0.19	D: 15 (18.75)	-0.13
4	1.03 (0.87, 1.13)	0.33	-0.89	44 (55.00)	0.33	A: 32 (40.00)	-0.28	B: 3 (3.75)	-0.10	D: 1 (1.25)	-0.09
5	0.92 (0.59, 1.41)	0.51	1.19	12 (15.00)	0.51	A: 38 (47.50)	-0.22	B: 18 (22.50)	0.05	C: 12 (15.00)	-0.27
6	0.94 (0.70, 1.30)	0.45	0.74	17 (21.25)	0.45	A: 26 (32.50)	-0.11	B: 18 (22.50)	-0.04	C: 19 (23.75)	-0.27
7	0.97 (0.56, 1.44)	0.38	1.30	11 (13.75)	0.38	B: 5 (6.25)	-0.07	C: 42 (52.50)	-0.16	D: 22 (27.50)	-0.07
8	0.95 (0.85, 1.15)	0.50	-0.11	30 (37.50)	0.50	A: 26 (32.50)	-0.22	B: 11 (13.75)	-0.08	D: 13 (16.25)	-0.30
9	1.03 (0.67, 1.33)	0.23	-2.16	64 (80.00)	0.23	A: 5 (6.25)	-0.00	C: 2 (2.50)	-0.20	D: 9 (11.25)	-0.19
10	1.01 (0.86, 1.14)	0.43	-0.28	33 (41.25)	0.43	B: 11 (13.75)	-0.25	C: 21 (26.25)	-0.31	D: 15 (18.75)	0.03
11	0.92 (0.24, 1.76)	0.52	2.20	5 (6.25)	0.52	A: 28 (35.00)	-0.09	C: 15 (18.75)	0.06	D: 32 (40.00)	-0.21
12	1.04 (0.44, 1.56)	0.09	1.67	8 (10.00)	0.09	A: 19 (23.75)	0.18	C: 35 (43.75)	0.01	D: 18 (22.50)	-0.25
13	0.96 (0.44, 1.56)	0.37	1.67	8 (10.00)	0.37	A: 43 (53.75)	-0.08	B: 6 (7.50)	-0.08	D: 22 (27.50)	-0.10
14	1.03 (0.80, 1.20)	0.35	-1.52	55 (68.75)	0.35	B: 10 (12.50)	-0.45	C: 14 (17.50)	-0.05	D: 1 (1.25)	0.06
15	1.10 (0.87, 1.13)	0.14	-0.34	34 (42.50)	0.14	A: 7 (8.75)	-0.38	B: 25 (31.25)	0.17	D: 14 (17.50)	-0.11

Wu and Adams state that an item is "less discriminating than the model predicts" if the fit value is greater than 1 (2007, p. 66), and if the MNSQ is substantially greater than 1, the item is advised to be removed.

From the table above, it can be seen that not all items have fit values that are closed to 1.00. Roughly half of them are below 1.00 and some others are over the expected value. At least, 3 items appeared to be quite farther from the expected value. This suggests that the test requires necessary revision in order to establish adequate construct validity to allow a sufficient degree of interpretation and use of the results of the test.

3.3 Discrimination Ability of Individual Item

For MC tests, it is essential that each individual item can discriminate students at their performance autonomy level. That means, students with higher ability will likely score higher while others with lower ability will tend to score lower in the test.

Discrimination index is the value that shows the degree of the discrimination ability of each individual item in discriminating students based on their performance autonomy level. It is assumed that higher ability students will likely select the correct answer while the lower ones will tend to choose a distractor. This is shown by the exact alignment between the value of the discrimination ability and the value of the point-biserial of those selecting the correct option. That means, the value of the discrimination index of each individual item is exactly the same as the value of the point-biserial of those selecting the correct answer. In other words, it could be assumed that the value of the point-biserial number of those selecting the correct option is used as the reference for generating the value of the discrimination ability of each individual item. In this regard, if the value of the point-biserial of those selecting the correct option is not the highest value of all the four options (in this test), the discrimination ability is argued to be problematic.

Wu and Adams (2007, p. 64) state that the value considered as high discrimination value is 0.4 and above, while the acceptable discrimination value is 0.2. Values lower than these two are hardly acceptable as they unlikely possess the ability of discriminating students according to their performance level. For this test, there are six individual items that have high discrimination index value –items number 1, 5, 6, 8, 10, and 11. Whereas, there are seven individual items that have acceptable discrimination value–items number 2, 3, 4, 9, 13 and 14. A few of these items actually have values close to the high value.

Apart from all the acceptable and high discrimination index values, there are two items that are problematic. Items number 12 and 15 have values lower than 0.2. Based on the analysis presented in the figure above, it could be found that the value of the point-biserial of those selecting the correct option is not the highest point-biserial values of all the given four options. In the case of item number 12, the highest point-biserial value is that of those selecting the option A. That means, higher ability students, assumed in the analysis, selected option A, rather than B as the correct answer. While for item number 15, the highest point-biserial value is that of those choosing option B, rather than C as the correct answer.

3.4 The Degree of the Plausibility of the Distractors

It is widely reported that providing plausible distractors is a challenge. Hughes (2012) and Weir (1990) reveal that good options that could attract test takers to select is often not available. One obvious result is that the provision of the distractors is pointless. Therefore, an analysis of all the distractors is essential when piloting a test in order to be able to finalise the test with better constructed distractors that are more attractive to as well as plausible for test takers (Fulcher, 2010, p. 173).

At this stage, it is important to point out that reconstruction of the distractors of half of all items is needed. These distractors had insufficient distracting ability and some were implausible. This echoes the fact that finding sufficiently diverting distractors for multiple-choice test items is evidently challenging.

It is also discovered that the degree of plausibility of the distractors does not necessarily determine the degree of discrimination ability of an item. The item number 12 and item number 15 have all plausible distractors. Each of the distractors attracted significant number of students but could not discriminate the students based on their ability. On the other hand, item number 4 has a near high discrimination value of 0.33, although it has one distractor, option D, which attracted only 1 student. Similarly, item number 14 even has slightly higher discrimination value of 0.35, although it has one deficient distractor.

4. CONCLUSION

In conclusion, at this stage, it is clear that the multiple-choice test, though developed by a teacher with strong credentials, was found to be somewhat problematic. Generally, the test requires some revision to enhance construct validity and has some items that exhibit high discriminability. Some items possess acceptable discriminability, but at least 13.5% of all the items with deficient discriminability. In terms of distracting ability, there were at least roughly half of all the items which had poor distractors, not to mention some that did not distract the test takers at all. Apart from all this quality, it was also discovered that the reliability of the test was far below the acceptable cut-off score for a cognitive test. These findings soundly echo what many scholars have highlighted that good multiple-choice tests are difficult to make.

Based on the findings, at least there are two implications that are worth highlighting. Firstly, teachers need to receive a proper training on assessment. Teachers need to know the importance of conducting assessment, what appropriate techniques and instruments to select and which ones would generate accurate information of the learners' learning progress and achievements. In the study, had the teacher been more aware of alternative objective instruments, he might probably have not decided to develop a multiple-choice test as his device to monitor his students' genuine learning progress.

Secondly, this study also suggests the importance of conducting wider-scale research. This is important to find out to what extent, classroom assessments in general are credible enough to inform teachers and users of the assessments about students' genuine level of achievement. If an assessment instrument is not credible, then the result of the test is deceptive. It would then be difficult to provide proper treatments, to deliver proper instructions, and it is difficult to make proper decision regarding the result of the assessment. Therefore, an attempt to carry out wider scale research could inform very important decisions in an attempt to improve the quality of the education in Indonesia.

REFERENCES

Bachman, L. F., & Palmer, A. S. (1996). *Designing language test.* Oxford: Oxford University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Pearson Education Inc.

Fulcher, G., & Davidson, F. (2007). *Language, testing, and assessment.* New York: Routledge.

Fulcher, G. (2010). *Practical language testing.* London: Hodder Education.

Hughes, A. (2012). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. Linn (Eds.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Miller, M. D., Linn, L. R., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). New Jersey: Pearson Education.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350-353.

Wu, M. & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Weir, C. (1990). *Communicative language testing.* New York: Prentice Hall.

Weir, C. (2005). *Language testing and validation: An evidence-based approach.* New York: Palgrave Macmillan.

Yulia, Y. (2014). An evaluation of English language teaching programs in Indonesian junior high schools in Yogyakarta province. (Unpublished doctoral thesis). RMIT University, Melbourne.

Zulfikar, T. (2009). The making of Indonesian education: An overview on empowering Indonesian teachers. *Journal of Indonesian Social Sciences and Humanities, 2,* 13-39.