Item analysis and item difficulty of multiple choice test of biology departments at Universitas Terbuka, Indonesia.

Diki Diki, Eko Yuliastuti,

Abstract

Abstract

The paper discusses a discrepancy of test items difficulty level between the test developers and students’ perception. Previous studies showed difficulty level was critical in multiple choice question test (Naqvi, Hashmi, & Hussain, 2010; Sim & Rasiah, 2006). High number of not valid test items was also reducing effectiveness of a test (Ratnaningsih & Isfarudi, 2010). The aim of the study is to compare the difficulty level of the test items according to the test developers and the difficulty level based on item analysis. The hypothesis is that there is a gap between two kinds of difficulty level, the effectiveness of the test item is low. The study used data from three examination result of BIOL4110 General Biology test at Universitas Terbuka, Indonesia of three consecutive semesters between 2014 and 2015. Participants of each examination were 469, 536, and 520 students. The passing rate of the test was 75,98%, 62,60%, and 73,09% consecutively. Data about difficulty level according to the students was gained from item analysis. Analysis of a relationship between difficulty according to test developers and according to item analysis is using paired sample t test. The analysis showed that in each semester, there is no significant differences of difficulty level between test developer judgment and item analysis result.

**Introduction.**

Universitas Terbuka is the primary distance education provider at post secondary level in Indonesia. Nowadays, it has more than 300,000 registered students from all over the country with 17,000 islands. The distance education program allows students to study regardless of geographical barriers, family and job responsibilities, and age (Holmberg 2005; Moore and Kearsley 2012; Zuhairi et al. 2009).

A department of Universitas Terbuka is Department of biology in the Faculty of Mathematics and Natural Sciences. The department was established in 1998. In 2016, there are more than 500 students in this department.

A significant problem of the department of biology at Universitas Terbuka is low test scores. In 2014, 29.14 % of the general biology students who took the final exam received very low scores and failed the examination (UT, personal communication).

One possible cause of the students failure during examination is accuracy of examination material difficulty level. Less accuracy of difficulty level caused students’ low level of success during examination. Examination material should have accurate difficulty level. It allows developing test material with appropriate balance of difficulty level in the examination material (Sim & Rasiah, 2006). However, there is lack of study in this subject at the department of biology at Universitas Terbuka.

Another suggested possibility of the studeents failure is high level of not valid test items. According to Ratnaningsih &Isfarudi (2010), there were findings of not valid test items in Universitas Terbuka. However, there should be more studies in the department of biology, since the study of Ratnaningsih & Isfarudi did not include biology students.

**Research question:**

* To what extent, level of difficulty predicts test score of biology students at UT ?
* To what extent is the match between item test difficulty level according to the test developer and students exam score.

**Hypothesis:**

There should be a correspondence between difficulty level of test developers and students perception about item test difficulty level.

**Theoretical framework:**

Ideal test should distinguish high achiever and low achiever test takers. One criteria of the ideal test is difficulty level. However, recent study showed that test items in UT still allowed test takers to do guessing during the test (Ratnaningsih & Isfarudi, 2013). Therefore, there should be a study about the level of difficulty in developing the test items. In order to measure difficulty level, the study used difficulty index (Naqvi, Hashmi, & Hussain, 2010).

Item difficulty is the percentage of correct answers to the total test items. The greater the index, the more difficult the test items are (Hotiu, 2006; Sim & Rasiah, 2006).

**Methodology**

The researchers will acquire the test scores of BIOL4110 General Biology. Then, we did item analysis based on Item Response Theory (Kubinger & Gottshald, 2007; Ratnaningsih & Isfarudi, 2013). Determination of level of difficulty used difficulty index (Naqvi, Hashmi, & Hussain, 2010). Analysis of a relationship between difficulty according to test developer and according to students answer is using paired sample t test.

The study used examination result three consecutive semesters in 2014 and 2015. Participants of each examination were 469, 536, and 520 students. The students who took the examination included students from other departements in the Faculty of Mathematics and Natural Sciences, such as statistics, mathematics, agribusiness, and food technology. The examination was conducted in a paper and pencil test in various location throughout Indonesia.

**Result**

The passing rate of the test was 75,98%, 62,60%, and 73,09% consecutively. Data about difficulty level according to the students was gained from item analysis. Analysis of a relationship between difficulty according to test developers and according to item analysis is using paired sample t test. The result was that t = -1.93 and sig t = 0.06. The analysis showed that in each semester, there is no significant differences of difficulty level between test developer judgment and item analysis result.

The analysis showed that none of the examination within three consecutive semesters showed relationship between difficulty level according to the test developer and according to the students answer. This result was different from previous study of Naqvi, Hashmi & Hussain (2011) in biology high school students in Pakistan, that showed low

**Conclusion**

As the paired sample t test did not show any differences of the difficulty level of the test developer and students, the test was considered good. The test developer judgement about level of difficulty match the students expectation.

However, there is a need for further research to search other causes of students low scores. In addition, the sample was only three consecutive semesters. A comparison study with other subject in other departments may revealed other possibilities.

References:

Holmberg, B. *Theory and Practice of Distance Education.* 2nd ed. New York: Routledge, 2005.

Hotiu, A. (2006). The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course.

Kubinger, K. D., & Gottshald, J. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats – An experiment in fundamental research on psychological assessment.

Moore, M. G., and G. Kearsley. (2012). *Distance Education: A System View of Online Learning.* 3rd ed. Belmont, CA: Wadsworth.

Naqvi, S. I., Hashmi, M. A. & Hussain, A. (2010). Validation of objective-type test in biology at secondary school level. Procedia Social and Behavioral Sciences 2. 20-31

Ratnaningsih J.D & Isfarudi (2013). Analisis butir tes obyektif ujian akhir semester mahasiswa Universitas Terbuka berdasarkan teori tes modern.

Sim, S. & Rasiah, R. I. (2006). Relationship Between Item Difficulty and Discrimination Indices in True/FalseType Multiple Choice Questions of a Para-clinical Multidisciplinary Paper.

Moore and Kearsley 2012;

Zuhairi et al. 2009