ӘОЖ: 378.046.2 DOI: 10.47751/skspu-1937-0024

IMPORTANCE OF ESTABLISHING VALIDITY AND RELIABILITY OF READING TEST

AKEBAI Berikbol
PhD Candidate Ankara University
Ankara/Turkish Republic
e-mail: berikbol0727@mail.ru

ERDENEBAATAR Chimedlkham
PhD, School of Foreign Languages
Mongolian University of Science and Technology
Ulaanbaatar/Mongolia
e-mail: cheeme99@gmail.com

Abstract. *This study presents the processes of developing and establishing reliability and validity of a reading test. In this respect, the study was conducted among 43 undergaduate students at Mongolian University of Science and Technology. Such detailed assessment is highly recommended for researchers who are in need of preparing pre and post tests which are different from each other. The results of students' achievement in this test were utilized to determine the quality of each particular item in terms of item difficulty and item discrimination analysis. Item difficulty, commonly known as p-value refers to the proportion of examinees that responded to the item correctly. To test the reliability of the reading test, item analysis was employed in terms of item difficulty, discrimination, average and variance of the test scores. The quality of the item as a whole indicates a reliable value Kuder-Richardson 20 (KR20) value of 0.71.*

Keywords: reading, Assessment of knowledge, reliability, validity, item analysis, item difficulty.

Introduction.

This article might be helpful to identify the differences among three confusing terminologies of 'assessment', 'evaluation', and 'testing'. In academic situations, particularly those college and university students are required to read a variety of authentic English materials such as textbooks, magazines, newspapers, journals, papers and so on. Students are also asked to read electronic books and other online materials from the Internet in order to gather information and broaden their knowledge.

Assessing student reading is a vital component of the teaching process. Besides, many studies and research projects have been conducted to examine validity and reliability of tests (Flippo & Schumm, 2009). Olson (2003) states "the keen interest of teachers in the dilemmas of testing has given rise to a movement toward exploring new forms of assessment, evaluation, grading, and reporting student progress, particularly in

the areas of reading" (p.323). According to Olson (2003), the term "assessment" refers to the deliberate use of many methods to gather evidence that the reader or writer is meeting his learning goals. As assessment is an ongoing part of instruction, it goes beyond response to offer feedback to both students and teachers about how reading is transmuting or how the learner is progressing. Similarly, Valencia (1990) defined assessment as a continuous and ongoing process. By observing and collecting information continuously, teachers can send a message to students, parents, and administrators that learning is never completed; instead, it is developing, and changing. According to McMillan (2004), assessment refers to the entire process of measurement, evaluation, and finally, use of the information by teachers and students. As identified by Noda (2003), assessment requires administering examinations to learn about the students' performances along with observing them in the classroom activities; however, evaluation has nothing to do with formal examinations since it deals with the students' performances in the classroom during the activities. On the other hand, testing requires administering specifically prepared examinations and is not interested in students' performances in the activities.

Regarding these definitions, assessment is a process that teachers engage in to determine what students know and are able to know and its rich data can inform and provide feedback about how to improve achievement and can be used to construct the criteria or benchmark for evaluation.

Reliability

Noda (2003) indicates reliability as a crucial element of standardized testing and points out that test-taker receive almost the same mark when they are delivered a reliable test for multiple times. This implies that if a reading test is reliable then the tester is sure that the test is consistent and test-takers perform almost the same at all times the test is delivered. If a group of test-takers perform much better or much worse in any test when compared with their previous scores on similar tests, then such a test cannot be regarded as reliable. The most common ways of assessing reliability is measuring 'stability or test-retest', 'alternate form' (Kaplan & Saccuzzo, 2001), 'internal consistency – Alpha' (Aiken, 2003), and 'interrater reliability or interrater objectivity' (Goodwin, 2001). To provide reliability, test-takers are required to use test techniques which are familiar to the test-takers; otherwise failure may occur as a result of unfamiliarity with the question types which results in an unreliable test.

Methodology

Participants of the study

Usually, it is too costly and time-consuming to collect data for all members of an actual population of interest, and therefore researchers usually collect data for a relatively small sample and use the result from that sample to make inferences about attitudes in a larger population (Warner, 2008, p. 3). In order to choose the participants for the study, the researcher used convenience sampling. A convenience sampling

consists of participants who are readily available to the researcher (Warner, 2008). Total 43 students participated in the study from the Department of Civil Engineering and Architecture who study in the academic year of 2018-2019.

Instrument

Reading part of Cambridge Preliminary English Test (PET) was conducted to check reliability scores of reading comprehension test for the study and to determine whether the reading comprehension test and the reading texts for the study would be appropriate in length, degree of difficulty, and content. The PET reading test was prepared by Cambridge ESOL Examination. PET is an English exam at intermediate level and reading texts are prepared for the level at B1 in the Common European Framework of Reference for Languages (CEFR). It consists of 3 parts which include in total 20 questions. CEFR guideline was used to describe achievements of the participants.

Instructional Assessment Resources (IAR 2011) believes that "an item analysis involves many statistics that can provide useful information for improving the quality and accuracy of multiple-choice. The quality of each item was analyzed to evaluate the quality of each item in terms of item difficulty and item discrimination. Item difficulty is basically the proportion of students who responded correctly to an item. In the meantime, item discrimination is a measurement to differentiate between the performance of students in the high score group and those in the low score group.

Result

Item Difficulty

The results of students' achievement in this test were utilized to determine the quality of each particular item in terms of item difficulty and item discrimination analysis. Item difficulty, commonly known as *p*-value refers to the proportion of examinees that responded to the item correctly.

To administer item analysis process, first the participants' answers were marked by the researcher and formulized on the Excel spreadsheet. Each correct answer was given one point and zero for each wrong answer. As a result, the average of difficulty was 80.7%. (Table 1)

To calculate *item difficulty* the number of all answers were added and sum was divided by total test-takers. (Table 1)

Table 1

Item Analysis of the Reading Test		
Items	(p) Item Difficulty	(r) Item Discrimination
Item 1	1	0.00
Item 2	87	0.31
Item 3	0.93	0.15

Item 4	1	0.00
Item 5	0.61	0.08
Item 6	0.77	0.54
Item 7	0.63	0.54
Item 8	0.79	0.46
Item 9	0.87	0.38
Item 10	0.75	0.38
Item 11	0.87	0.46
Item 12	0.83	0.31
Item 13	0.81	0.38
Item 14	0.81	0.46
Item 15	0.46	0.69
Item 16	0.69	0.54
Item 17	0.87	0.15
Item 18	0.71	0.54
Item 19	0.85	0.38
Item 20	0.91	0.31

Item discrimination

This paper stresses for utilization of discrimination coefficient considering 43 students with the intention that every single person's performance was taken into consideration. The discrimination coefficient, the Pearson *r,* for each item was computed using Statistical Package for the Social Sciences (SPSS) version 23 The Pearson, *r* coefficient ranges between -1 and 1. Parallel to the discrimination index, a higher value indicates a powerful discrimination power of the respective test. A highly discriminating item reveals that students with high score got the item right and students with low score answer the item incorrectly. Items with negative values should be rejected for the reason that negative value reflects the opposite effects of discriminating power for that particular item.

Reliability of the reading test.

The reliability was computed in Kuder and Richardson Formula 20 and Cronbach's alpha.

Kuder-Richardson 20, a formula which is based on item difficulty was used to analyse internal consistency of section A in the string instrument comprehensive test. The value of KR20 range between 0 to 1. The closer the value to 1 the better the internal consistency. The KR20 formula is commonly used to measure the reliability of achievement test with dichotomous choices. According to Fraenkel and Wallen, researcher should attempt to generate a KR20 reliability coefficient of .70 and above to acquire reliable score.

To test the reliability of the reading test, item analysis was employed in terms of item difficulty, discrimination, average and variance of the test scores. The quality of the item as a whole indicates a reliable value Kuder-Richardson 20 (KR20) value of 0.71. (Table 2)

Table 2 provides the results obtained from the analysis of student comprehensive test score.

Table 2

Descriptive Statistics	
N (total number of students)	*49*
Mean	*80.7*
Standard Deviation	*2.89*

Reliability analysis revealed a Cronbach's alpha score $\alpha = 0.710$ over 20 items in the reading test.

Conclusion

This paper includes information about establishing the reliability and validity of a reading test, as well as a description of the development procedure of the test. After such detailed validity and reliability analyses, it might be possible to report about a reading test's restrictions, such as readability of the texts, what grades the test is appropriate for, and the how discriminative the questions in the test are. The study aimed at describing the process of establishing validity and reliability of a reading test in detail with the intention of providing valuable information about multiple assessment criteria both to teachers of reading who rely on reading tests to determine reading skills of their students and researchers who are in need of reliable reading assessment tools for their pre and post tests. Establishing such validity and reliability analyses might also be beneficial for testers as they depend on assessment tools for making decisions about the candidates.

In order to offer any opinions about the quality of a reading test, some assessment criteria are supposed to be administered. Assessing any reading test with just a single criterion may not hinder realistic results. Therefore, evaluating reading tests in terms of multiple factors may assist teachers, researchers, and testers to decide for themselves which reading test is most appropriate for their particular needs. The general tendency to assess a reading test is dealing with its validity and also reliability. Such an assessment requires reading tests which are free of bias and distortion. However, such analyses do not necessarily reveal exact difficulty of the texts in the test as reliability focuses on question items rather than the texts in the test. In addition , calculating readability also gives an idea about the difficulty of a text. Nevertheless, readability analyses can also be considered superficial as they merely deal with either word or sentence lengths. Therefore, vocabulary frequency analysis may assist testers to assess their texts more deeply.

Implications

Such detailed assessment of a reading test in terms of its validity and reliability is highly recommended for researchers who are in need of preparing pre and post tests for experimental studies. Then, they will be able to administer pre and post tests which are both different from and identical to each other. However, it might be very tiring for reading teachers to administer such detailed analysis for their reading tests. Due to their profession, researchers might be aware of the importance of establishing validity and reliability for their reading tests; however, this may not be the case for teachers as their principal goal is teaching rather than researching. Nevertheless, teachers should also be encouraged to use valid and reliable tests to assess their students' reading skills. It might be beneficial to assist reading teachers at any grade to achieve this goal.

References

Aiken, L. R. (2003). *Psychological testing and assessment* (11th ed.). Boston: Allyn and Bacon.

Flippo, R.F., & Schumm, J. (2009). Reading test. *Handbook of College Reading and Study Strategy Research.* 408-464.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science*, 5, 13-14.

Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principle, applications and issues* (5th ed.). Belmont, CA: Wadsworth

McMillan, J. (2004). *Educational research: Fundamentals for the consumer*. New York: Longman.

Instructional Assessment Resources. (2011). *Item Analysis. Retrieved November 9, 2013 from University of Texas at Austin.*

Noda, M. (2003). Evaluation in reading. In H. Nara & M. Noda (Eds.), Acts of reading: Exploring connections in pedagogy of Japanese (pp. 197-222). Honolulu: University of Hawai'i Press.

Olson.C.B. (2003). *The reading/writing connection: Strategies for teaching and learning in the secondary Classsroom*. New York: Allin &Bacon/ Longman Printing.

Valencia, S. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *The Reading Teacher*, 43, 338-340.

Warner, R. (2008). *Applied Statistics*. London: Sage Publication.

Оқу тестінің негізділігі мен сенімділігін орнатудың маңыздылығы

ӘКЕБАЙ Берікбол
Анкара университетінің PhD докторанты
Анкара/Түркия Республикасы

ЕРДЕНАБААТАР Чимэдлхам
PhD докторы, Моңғолия Ғылым және Технология

Университеті «Шетел тілдері» кафедрасы
Ұланбатыр/Моңғолия

Аңдатпа. *Бұл зерттеуде оқу тестінің сенімділігі мен дұрыстығын құру және бекіту процестері келтірілген. Осыған байланысты зерттеу Моңғолия Ғылым және Технология Университетінің 43 магистранты арасында жүргізілді. Мұндай егжей-тегжейлі бағалау бір-бірінен ерекшеленетін алдын-ала және кейінгі тестілерді дайындауды өте қажет ететін зерттеушілерге ұсынылады. Студенттердің осы сынақтағы жетістіктерінің нәтижелері әр нақты элементтің сапасын элементтердің қиындықтары мен элементтерді кемсітуді талдау тұрғысынан анықтау үшін қолданылды. Әдетте р-мәні деп аталатын заттың қиындығы зерттелушілердің дұрыс жауап берген үлесін білдіреді. Оқу тестінің сенімділігін тексеру үшін элементті талдау элементтердің қиындықтары, кемсітушілік, тестілеу баллдарының орташа және дисперсиясы тұрғысынан қолданылды. Тұтастай алғанда заттың сапасы Kuder-Richardson 20 (KR20) 0,71 сенімді мәнін көрсетеді.*

Кілт сөздер: *Оқу, білімді бағалау, сенімділік, жарамдылық, затты талдау, заттың қиындығы.*

Важность установления действительности и надежности считывания теста

АКЕБАЙ Берикбол
PhD докторанты университета Анкары
Анкара/Турция

ЕРДЕНАБААТАР Чимэдлхам
PhD, Кафедра иностранных языков
Монгольский университет науки и технологий
Уланбатыр/Монголия

Аннотация. В данном исследовании представлены результаты разработки и подтверждения надежности и валидности теста по чтению. В связи с этим исследование проводилось среди 43 студентов старших курсов Монгольского университета науки и технологий. Такая подробная оценка настоятельно рекомендуется для исследователей, которым необходимо подготовить предварительные и последующие тесты, которые отличаются друг от друга. Результаты успеваемости учащихся в этом тесте использовались для определения качества каждого конкретного задания с точки зрения его сложности и анализа различения заданий. Сложность задания, обычно известная как p-значение, относится к доле испытуемых, которые правильно ответили на задание. Чтобы проверить надежность теста чтения, был использован анализ заданий с точки зрения сложности заданий, различения, среднего и дисперсии результатов теста. Качество товара в целом свидетельствует о достоверном значении Kuder-Richardson 20 (KR20), равном 0,71.

Ключевые слова: чтение, оценка знаний,надежность, валидность, анализ заданий, сложность задания.