Assessing Reading and Vocabulary

Amber Kantner, Hannah M. Landers, & Matteo Musumeci

Northern Arizona University

Abstract

Creating a test that assesses the needs of the students along with meeting the needs of teachers, administrators, and stakeholders can be a difficult process. While the journey to create a reliable and valid test can trying, it is important to make sure that any product created meets the needs of all parties involved. This final report attempts to address the whole test creation process based on trial and error, from the initial needs assessment all the way to analyzing the reliability and validity of a test. The test creators decided on constructing a high-intermediate reading and vocabulary test for an English for academic purposes (EAP) class and then measured whether or not the test measure what it was supposed to. This final report will cover the description of the test, the methods employed, an analysis of the test results, and a discussion. By the end of the report, readers should have a better understanding of the test creation process and how to interpret and analyze the results.

Background

Assessing reading ability can be extremely difficult as it is an internal process and many testing techniques overlap with other skills that are not necessarily being assessed. Nonetheless, the test created aimed at examining students' reading comprehension and their understanding of necessary vocabulary while attempting to incorporate as little other skill areas as possible. The test that was created aimed at assessing students' literal understanding and inferencing ability when reading a text passage. While there is no conclusive study that shows authenticity of a test passage aids in test taker performance (Lewkowicz, 2000), the test attempted to use authentic materials that students may encounter in real life and for research, additionally the questions about the text were intended to assess their ability to extrapolate pertinent information.

The test created was created for and administered to three Level 4 (high-intermediate) Reading and Writing classes in Northern Arizona University's (NAU) Program in Intensive English (PIE). In creating a test, there are several aspects that are required. This paper will delve into all the requirements including (a) a description of the test that includes important facets of test creation, (b) the methods which includes how the test was actually administered, (c) the results that include statistics, reliability, and validity, and then move into (d) a discussion of the results. This final report will wrap up the process of creating a test and suggest improvements for creating a more accurate test in the future.

Research Questions

When creating this project, the test designers had a few question in mind. Can we create a viable measurement tool to test students' abilities in reading and vocabulary? Additionally, can we interpret conventions for creating tests to successfully measure ability? Lastly, can we use the prescribed tools to assess whether our test was successful?

Methods

There were 46 Level 4 students from three different Reading and Writing classes at the PIE who participated in our test (4A n = 15; 4B n = 14; 4C n = 17). It was decided to test all three sections of Level 4 in order to have as many data points as possible to enable more precise statistical data derived from the results. The classes were composed of both male and female native Arabic speakers, Portuguese speakers, and Chinese speakers who are studying in the United States on F1 visas. The students were familiar with American-style testing procedures and were familiar with this particular test style as it is based on their classroom activities.

The test started with a cover page that informed students that the test had two reading passages along with different types of questions associated with each. It also tells students that they will have 50 minutes to complete the test. The first passage talked about compulsive gambling. Students were then asked to answer eight multiple choice questions asking about main ideas, details, inferences, and vocabulary, they were then asked to answer two questions (the first a main idea question; the second a detail question) about the text, and lastly they were asked to produce two sentences using different vocabulary questions. The second passage was about people's addiction to natural resources. The questions followed a similar format, only for this passage, the two supply questions about the text were a detail and an inference.

This test was administered during week 13 of the program. The test took approximately 60 minutes including instructions and was administered only one time. The teachers along with one test creator were the administrators. Each teacher received a proctoring guide (see appendix F) to help facilitate giving the test. The proctoring guide informed test givers of the required materials, the classroom preparation, how to actually give the test, and test collection.

Results

Each test item was evaluated for its *P* and *D* value (see Table 1). *P* is also known as the item discrimination which is how the test creator can evaluate item difficulty (Miller et al., 2009). Items with a high *P* value are considered too easy as that means the majority of students knew the correct answer. All of the *P* values for this test were 0.74 and above. The next category is the *D* value, or the item discriminating power shows if students are performing as predicted; in other words, students in the upper group of test takers should be performing better (Miller, Linn, & Gronlund, 2009). Any items that have a *D* value of zero or a negative show no correlation thus are not accurately representing actual student understanding. The majority of the *D* values are low; very few rise above a value of 0.30.

The chart to represent the test's descriptive statistics was first separated into several sections (see Table 1). For each of these items we looked at several variables. For each section the number of items was listed as K. The minimum (Min) recorded scores and the maximum (Max) recorded scores were listed; in this section there was no one that missed all of one type of question nor was there any section where there was not at least one perfect score. The mean (Mean) of the scores for each section of the test were recorded and in most cases, the mean was above the determined cut score of 70%. The standard deviation (SD) shows how far away the score is from the mean. In most cases, the values fell within one standard deviation of the mean. The major exception is for the test overall where the value was 2.59 standard deviations from the mean.

Table 1

Descriptive Statistics and Reliability

	K	Min	Max	Mean	*SD*	*r*	ρ_0	SEM
Total	24	13	24	20.28	2.59	0.62	0.29	1.60
Main Idea	6	3	6	5.33	0.87	0.32	0.15	0.72
Detail	5	1	5	3.74	0.91	0.27	0.19	0.78
MI & DET	11	5	11	9.07	1.34	0.67	0.42	0.86
Inference	5	3	5	4.40	0.68	-0.11	N/A	0.72
Vocabulary	8	3	8	6.83	1.34	0.51	0.28	0.94

Note. N= 46; K= number of items; Min = minimum score; Max = maximum score; SD = standard deviation; r = Cronbach's alpha; $\rho 0$ = kappa coefficient; SEM = standard error of measurement

The sections for Cronbach's alpha (r), the kappa coefficient (ρ_0), and the standard error of measurement (SEM) require a little more background information to understand. Cronbach's alpha describes how well a set of items fit in with the overall test. The value should be between 0.60 and 0.80 for teacher made tests. The kappa coefficient which is a measure of consistency and used to measure the reliability of the test (Subkoviak, 1988). Subkoviak (1988) suggests that classroom tests should have a value for the kappa coefficient between 0.35-0.50 in order to show that the test is measuring students' ability accurately. Last of all is the SEM. The SEM is also a measure of reliability and "can be used to in defining limits around the observed score within which one would be reasonably sure to find the true score" (Doppelt, 1956, p. 1).

Relevance to PIE and Second Language Learning

When any skill is being taught, it is necessary to assess the progress of the students attempting to master that skill. It is vital that all language teachers are able to assess students' learning in order to promote student success and for teachers to improve their teaching techniques to aid in student improvement. Researching in the PIE means that the test designers were able to improve their skill set and promote learning by means of assessment.

References

Doppelt, J. E. (1956). *How accurate is a test score?* NY: The Psychological Corporation.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing 17*(1), 43-64.

Miller, D.M., Linn, R.L., & Gronlund, N.E. (2013). *Measurement and assessment in teaching* (11th ed.). Boston, MA: Longman Pearson.

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement 25*, 47-55.