Assessing Reading and Writing Skills for Classroom Purposes

Roman Lesnov, Panjanit Chaipuapae, & Meishan Chen

Northern Arizona University

Fall 2014

Abstract

This report provides the results of the research on creating an ESL formative test and its implementation into ESL classroom assessment. The main purpose of the study is to investigate the validity and reliability of the developed test as well as the effectiveness of its items. The report includes the description of the test, methods, materials, administration procedures, and the discussion of the test results. The interpretation of the results showed that the test appeared to be reasonably valid and moderately reliable though may require some revision and modification. This analysis yielded a number of improvements that researchers would have to make in order to increase the reliability of the test.

Assessing Reading and Writing Skills for Classroom Purposes

As language teachers, one of our main responsibilities is to provide meaningful instruction throughout the course. Needless to say, we would also want to know whether our instruction is effective and worthwhile and whether there are any learning difficulties during instruction so that we can accommodate students' learning process at best before the end of instruction. The purpose of the study is to develop a test given during instruction in order to provide the basis for formative assessment. In other words, we based our framework for developing the test on Miller, Linn, and Gronlund's (2013) statement: "They [tests and assessments] are used to monitor learning progress, detect misconceptions, encourage students to study, and provide feedback to students and teachers" (p. 141).

The study is intended to investigate the language areas of reading and writing in English. Alderson (2000) suggested that we should be careful with integrated testing (e.g., reading and writing skills) because one skill might contaminate the measurement of the other. However, the test items were carefully constructed according to the specifications of the test which are aligned with the course syllabus and classroom materials learned in class. In conducting the study, the researchers had an opportunity to be exposed to the real world tasks in testing and assessment, namely determining the purposes of measurement, developing specifications, selecting appropriate assessment tasks, preparing relevant assessment tasks, assembling the assessment, administering the assessment, appraising the assessment, using the results, and improving learning and instruction (Miller, Linn, & Gronlund, 2013, p. 140). One of the researchers, who was a teacher of the class, gained useful information from the results of the study.

Construct Definition

In general, the test is intended to measure academic reading and writing skills that students are expected to have acquired during instruction, as well as the knowledge of target vocabulary and the ability to productively use it in writing. So, the three subconstructs are included in the construct definition – academic reading, writing, and vocabulary. In turn, each of them has a set of specific abilities, which are in full compliance with classroom objectives.

Subconstruct 1. Academic Reading subconstruct incorporates reading simplified academic texts expeditiously and carefully, which can be assessed by measuring the abilities to:

- identify main ideas and major details in significantly simplified academic texts;
- make inferences related to significantly simplified academic texts;
- recognize textual organization patterns in significantly simplified academic texts.

Subconstruct 2. Vocabulary subconstruct incorporates receptive and productive vocabulary knowledge, which can be assessed by measuring the abilities to:

- recognize the meaning of target content-based vocabulary and meaningfully use it in writing.

Subconstruct 3. Academic writing subconstruct incorporates writing a three-paragraph compare-contrast essay, which can be assessed by measuring the abilities to:

- organize the structure of the essay:
    - write a good introduction;
    - write a good body of the essay;
    - write a good conclusion;
- compare and/or contrast two objects:
    - find three differences and/or similarities between two objects and express them clearly in written English;
    - write on topic and fulfil the requirements of the prompt;
    - effectively use connecting words and transitional devices;
- use grammatically and semantically accurate language.

Research Questions

Research Question 1. To what extent do the students learn the course content after 11 weeks' instruction?

Research Question 2. To what extent does the test provide a valid measure of the defined construct and its subconstructs?

Research Question 3. To what extent are the test items for Reading Comprehension appropriate? This question includes three sub-questions: To what extent are the test items appropriate in terms of (a) item difficulty, (b) item discrimination, and (c) plausibility of distracters?

Research Question 4. To what extent is the assessment reliable?

Methods

Participants

Data were collected during Fall 2014 with 15 international ESL students in the Program in Intensive English (PIE) at Northern Arizona University (NAU). The proficiency level of the students can be estimated as being equivalent to TOEFL score of 32-44 overall. The participants were students at Level 3 according to the PIE framework of English language proficiency. The group included two students from Brazil, three students from China, and 10 Arabic-speaking students. Informed consent was obtained from all participants in the study.

Administration

The test was administered to the students in a pencil-and-paper format during the 11th week of instruction. Before starting the test, it was the teacher's responsibility to explain the content of the test to students and to give exhaustive directions. In terms of the timing, students were allowed 75 minutes to complete the test. During this time, the students were not allowed to have a break or to leave the classroom. The test was administered in two classrooms and the

students were separated from each other as much as possible to reduce cheating. As expected, the teachers were monitoring the test-takers during the whole period of administration.

Results

With regard to the first research question, to what extent do the students learn the course content after 11 weeks' instruction, it was found that the distribution of the students' scores in the three subtests and their total scores are negatively skewed, as expected for a CRT test, indicating a progress in the understanding and use of the knowledge after 11 weeks' instruction.

In order to answer Research Question 2: To what extent does the test provide a valid measure of the defined construct and its subconstructs, the first thing to mention is the relevance of test items with regard to the defined subconstructs. The items were developed in accordance with instructional objectives and were closely related to the content of the course. The negative skewness of the results data may serve as the evidence that the items were construct-appropriate to elicit the expected students' performance on the test. In addition to that, as a check on construct and content validity, the test items were sent to PIE Reading and Writing course coordinator to obtain suggestions and modifications. This also enhanced the overall validity of the test.

The second piece of evidence lies in the reliability of scores, which is another aspect of construct validity. The reliability indexes for each of the subtests, which are 0.76 for Reading Comprehension (KR-20), 0.93 for Writing Vocabulary (inter-rater agreement coefficient), and 0.60 for Writing Compare-Contrast Essay (inter-rater agreement coefficient), demonstrate consistency in students' performance. This, in turn, validates the assessment.

Regarding the scores use and impact, the consequences of the test results have had intended outcomes. The scores were used to make interpretations about students' progress, which

was anticipated to be substantial. As a whole, the test showed the expected achievement of the students and success in teachers' instructional practices. In other words, the test results achieved intended consequences for teaching and learning, which added to the validity of the test.

Finally, the external relationship between this skills assessment test (SA) and the Achievement Test # 2 (AT) expressed by the correlation coefficient of 0.77 was also considered as a source of convergent predictive validity. It proved that the two tests measured similar constructs, which testifies to the SA construct validity.

Research question 3 asks to what extent the test items for Reading Comprehension are appropriate in terms of item facility (IF), B-Index, and plausibility of distracters. Overall, the item analysis for Reading comprehension part showed that most test items are relatively effective in reflecting the gain in this skill.

The distracter analysis would help us improve the multiple-choice items which are not effective as plausible distracters. Taken B-Index data into account, some items with B-Index of 0.33 showed that the items could not do well in discriminating student; therefore, the distracters should be revised accordingly. Overall, most distracters were appropriate.

In order to answer research question 4, to what extent is the assessment reliable, KR-20 and Cronbach's Alpha were calculated for the reliability of Reading Comprehension, and the agreement coefficients were estimated for the reliability of the Vocabulary scores and Compare and Contrast Writing scores. It was found that KR-20 equals to Cronbach's Alpha (KR-20=0.76, SEM=0.87) for the Reading Comprehension test. With regard to the reliability of the vocabulary and writing tests, the agreement coefficients were 0.93 and 0.6 respectively.

In order to examine the internal correlation of the test, correlation between the subtests and the total score was also calculated. A moderate correlation of 0.543 was found between the

writing score and the reading comprehension score. A strong relationship was found between the writing score and the total score, as well as between the reading comprehension score and the total score.

To ensure the validity of the test, the researchers investigated how well the students' total scores for this SA test correlated with those for the AT. The AT assessed the same three subconstructs: academic reading, vocabulary, and academic writing. The only difference was that the academic writing subconstruct in the AT included one more type of essay, namely summary writing. So, it was assumed that the students' SA test results would be a suitable predictor of the students' performance on the AT. The correlation analysis was conducted to justify this assumption and thus to bring the additional evidence to the construct validity of the SA.

To calculate the correlation index, the students' results for the AT were obtained from the PIE Assessment Department. Out of the 15 students one did not take the AT. That is why this student's results for the SA test were unheeded so that the researchers could have equal numbers of students for the sake of the correlation analysis. The calculated index between the total scores of the two tests was 0.77, which gives rise to the claim that the tests measured a similar constructs. This interpretation is supported by that fact that the two tests were very similar in format. As well, the degrees of standard deviation for the tests' total scores were almost equal and quite large (2.30/30 or 7.67% for the SA and 6.79/85.94 or 7.90% for the AT), which is a favorable condition for the unbiased interpretation of the correlation index.

In addition, correlation indexes between the corresponding SA and AT subconstructs were computed. The indexes (0.83 for Reading Comprehension, -0.24 for Writing Vocabulary, and 0.52 for Writing Compare-Contrast Essay, see Table 5) revealed that the Reading Comprehension subtests had the strongest relationship with each other. The negative correlation

between the results for the vocabulary sections can be partly explained by the fact that the different scoring systems were applied for rating.

Conclusion

The study provided an opportunity for the researchers to construct, administer, analyze, and interpreting the test results in the form of formative assessment. The validity and reliability of the test were also analyzed to make our interpretation and usefulness of the test valid and reliable. Even though the test needed some revision in terms of the test items and the grading rubrics, the developed test provided valuable information for both the students and the teacher during instruction.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Bachman, L. & Palmer, A. (2010). *Language assessment in practice*. New York: Oxford University Press.

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