Evaluation of Advanced Level Writing Assessment

Katherine Eccles

Northern Arizona University

Abstract

The purpose of this project is to determine the validity and reliability of a writing assessment designed for advanced level writing students. The test was designed for a classroom of only eleven students, which means it will not be valid to make generalizations based on this particular test's results by claiming reliability. The writing assessment was designed for level five students at the Program in Intensive English (PIE). Level five is the highest level at the PIE, so these ESL students are advanced writers who will enter a university the following semester if they pass this final level in the program. The writing assessment investigates innovative ways to test students' knowledge of academic writing by testing the students' ability to answer questions about writing, as well as measure their ability to write a cohesive process analysis essay. By testing students' declarative and procedural knowledge, the test will also reveal a possible correlation of these two types of knowledge, and may answer the question of whether or not it is important to test multiple types of knowledge related to the writing process.

Keywords: ESL, academic writing, procedural knowledge, declarative knowledge

Evaluation of Advanced Level Writing Assessment

Background

Previous research in the area of writing assessments informed the development of the advanced level wiring classroom assessment. For the procedural knowledge portion of the assessment, Ling He's (2012) study was considered. He conducted a study with fifty students studying at a Canadian collage whose levels ranged from basic to advanced. Each student wrote a response to a prompt requiring general knowledge, and wrote to a second prompt requiring specific knowledge. Through the study, He (2012) found that students from all levels preformed significantly better on the general topic than they did on the specific topic. Students' responses to the general topic had better content and idea development, organization, a clearer position, developed concluding paragraphs, and better language use. This study revealed the importance of using appropriate test prompts (He, 2012). Therefore, the background knowledge of the students in the advanced writing class at the PIE was an important factor in the process of choosing an appropriate writing prompt.

Research Questions

1. Is there a connection between student's declarative knowledge and procedural knowledge for the specific writing sub-construct of language use?

2. Does the writing test adequately measure the target construct and sub-constructs?

3. Is the writing test and scoring process reliable?

Methods

Ten advanced level writing students in the PIE participated in this study. All students graduated from high school in their home country, and their ages ranged from 18 to 21. Nine out of the ten students were Arabic and one student was Japanese. All students were enrolled in the semester-long writing level five course at the PIE, and the test was given about half way through the semester to assess their progress in meeting the learning objectives. The purpose of this advanced level writing class was to prepare students to enter an English-medium university. Therefore, the assessment was developed based on the learning objectives of the course which consisted partly of students being able to write a cohesive essay, use appropriate academic language in their writing, and use aspects of the writing process to improve their final written product.

The assessment had three main sections: 1) a timed argumentative essay, 2) eighteen items on the use of modal and transition words, and 3) an error correction task that had twelve language use errors for students to identify with an error code. The teacher explained all test instructions verbally as students read the instructions in their test packet. The teacher asked if students had any questions, and then was available during the test to answer individual questions. Students had one hour to complete the first section, which was the timed essay, and then they were given a ten-minute break. The table of specifications below (see Figure 1) displays the breakdown of the sub-constructs that were tested by the assessment in the horizontal rows, and the target sub-constructs in the vertical rows.

	Essay Content Band	Essay Organization Band	Essay Source Use Band	Essay Language Use Band	Transitions	Modals	Error Correction
Language Use				*X*	*X*	*X*	*X*
Content	*X*						
Organization		*X*					
Source Use			*X*				
Editing				*X*			*X*

Figure 1. Table of Specifications chart.

This assessment first examined student's procedural ability to write a process analysis essay in an hour. Students were given ten minutes to read the one-page source, ten minutes to fill out an outline for their essay (which was not graded), and 45 minutes to write and edit their essay. The rubric graded students specifically on content, organization, source use, and language use in their essay. Each sub-construct was rated on a scale ranging from one to five; five being the highest possible band score. The analytical rubric reflected the underlying construct definition, which encompassed many specific and discrete writing abilities to describe a more complete picture of a successful academic writer.

Then, students were tested on their declarative knowledge of grammatical structures and features of academic writing. Students were asked to answer four multiple-choice items in which they choose the most appropriate modal based on a description, and then they chose the correct modal from three different possible modals in a cloze-type activity. Students answered questions following the same structure to assess their knowledge of transition words. The third section of the test was designed to measure students' ability to identify and code errors in an essay. The types of errors present were chosen based on what had been taught in class and information about the number of errors present in the essay was given to the students: three subject verb agreement errors, three spelling errors, three word form errors, and three article errors. Students received one point for identifying the correct error, and one point for writing the correct error code next to the error.

Results

As shown in Table 1, the mean or average score was 27.4, which is equivalent to a 72% on the test as it had 38 total possible points. The median was about two points higher at 29.5. Because of the small number of test takers, the descriptive statistics are influenced by the two students who received a low score of 18. The spread of scores appears large, as the standard deviation is 5.25; however, the two low outliers heavily impact this high number. It is notable that the scores were still negatively skewed, meaning more scores are clustered around the higher end of the scale.

Table 1

Descriptive Statistics

Total Test Scores	
N	10
Mean	27.4
Mode	18
Median	29.5
Min	18.3
Max	32.5
Midpoint	25.4
Range	14.2
Variance (N<30)	27.56
SD (N<30)	5.25
Skewness	-1.15
Kurtosis	-0.05

To look more closely at the test, all the sub-sections were analyzed individually. The descriptive statistics for both the total scores of the essays, and the band scores of the essays can be interpreted by looking at Table 2. Similar to the total scores, there are two outliers present in the essay score results, which especially impact the mean and standard deviation of the essay scores. The reported mean for the total essay score was 14.96, whereas the mode was 16.75, and the median was 15.88. They both reflected higher values because they are not as influenced by the two low scores. Looking at the band scores on the chart below showed that the organization band had the highest mean (3.94), and the source use band had the lowest mean (3.46). The source use band also had the greatest spread of scores with a standard deviation of 1.32, while the rest of the bands had a standard deviation under 1.00. The total essay scores, and all the band scores except language use were negatively skewed.

Table 2

Descriptive Statistics for Argumentative Essay

	Total Scores	Content	Organization	Source Use	Language Use
N	10.00	10.00	10.00	10.00	10.00
Mean	14.96	3.84	3.94	3.46	3.66
Mode	16.75	4.00	4.00	4.00	4.00
Median	15.88	4.00	4.00	4.00	3.54
Min	10.33	2.38	3.25	1.00	3.25
Max	16.75	4.25	4.63	4.38	4.00
Midpoint	13.54	3.31	3.94	2.69	3.63
Range	6.42	1.88	1.38	3.38	0.75
Variance (N<30)	5.27	0.32	0.17	1.74	0.10
SD (N<30)	2.30	0.57	0.41	1.32	0.31
Skewness	-1.50	-2.26	-0.21	-1.64	0.05
Kurtosis	0.98	5.65	-0.31	1.09	-1.79

The second section of the test was focused on assessing student's declarative knowledge of both transition and modal use. The two sets of items were combined to calculate that the mean was 12.40, the mode was 15, and the standard deviation was 3.20. The last section of the assessment was the essay error correction section. The mean was 4.10, the mode was 5.00, and the median was 4.00. The standard deviation was 2.92 so there was spread in the scores even through overall scores were low. Not surprisingly, the scores were positively skewed further revealing the difficulty of this section.

Next, the test's internal consistency reliability was evaluated by looking at the KR-21 values for the transition and modal items along with the students' overall scores. Interestingly, the transition items showed the greatest reliability at 0.84. However, there were only 9 items, so the high value does not mean the test section was reliable. The modal item's KR-21 value was 0.51, so it was lower than the transition item scores. Lastly, the total score KR-21 value was 0.74. Even though this is a high value, it is not a valid representation of the test's reliability partly because non-objective test items were considered in the value.

Evidence of construct validity is shown through correlations between sections of the test that measured the same sub-construct. A correlation of 0.76 was identified between students' combined scores on the transition and modal questions, and their scores on the language use band of the essay as seen in the graph below. Similarly, the error correction and language use band had a correlation of 0.75. The highest correlation was seen between the total essay score, and the combined transition and modal section scores at 0.82. This is visible in the graph titled, "Essay and Language Use Correlation."

Relevance to the PIE and Second Language Learning

Designing and evaluating this test helped me to discover the importance of defining constructs of the skills I teach ESL students at the PIE. I also better understood the complexities of creating a valid and reliable test that measures only what is actually taught in class. I wanted to discover a relationship between the different sub-constructs of language use, and after completing this project I see that there is a possible connection between declarative and procedural knowledge of language use. However, because this was a small-scale assessment, generalization cannot be made, and further research would need to be conducted. I was interested to see how students' ability to correct errors in an essay was related to their scores on their essay, and I would like to study this relationship more extensively in the future. It was helpful as a teacher to see that students could not identify or code errors. As a result, I was able to design additional instruction and practice activities to help students notice errors. I further realized that accurate language use is an important skill to have going into the university, so it was worth providing extra practice in class. Overall, I noticed that the test helped me identify which students were meeting the course objectives, and how I needed to modify my instruction. I plan to continue to develop and analyze the assessments I develop throughout my teaching career in order to inform my instruction, and to help my students identify their ability to meet class objectives.

References

Celce-Murcia, M., Brinton, D. M., & Snow, M. A. (Eds.). (2014). Teaching English as a second or foreign language (4th ed.). Boston, MA: Heinle Cengage. 222-237.

Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing*. *18*(2), 207-224.

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing*, *29*(3), 443-464.

Miller, M. D., Linn, R., & Gronlund, N. (2013). Measurement and assessment in teaching. (11th ed.). New York, NY: Pearson.