Post-test Analysis to Increase Stakeholder Confidence: A Case Study for an English Language Assessment in Kurdistan

Mark Frohnsdorff

University of Kurdistan, Erbil, Iraq, Email: m.frohnsdorff@ukh.ac

Received: March 15, 2014 Accepted: April 12, 2014 Online Published: September 25, 2014

Abstract

This paper describes a case study where post-test analysis was used to critically evaluate specifications and items in light of student results. Unusually, this was done after the test was used live during a university semester. From the analysis, poorly performing items were removed from the test and a new overall total established. A new scoring system was implemented to accommodate this practice. The aim was to improve reliability and thus confidence in the scores of stakeholders, but primarily students. The absence of such a facility had previously affected the quality of testing on the University of Kurdistan-Hewler's Language Program, and critically undermined confidence in the assessment process. UKH witnessed this at the end of the 2012/2013 academic year. Some students rejected their score and decisions about their progression, resulting in student demonstrations reported in local media in Kurdistan.The author argues that simple steps can be taken to assure quality assessment. This process must involve all teaching staff, be transparent to students, and be context-sensitive. In conclusion, language assessment at UKH can now be measured quantitatively and qualitatively.

Keywords: Assessment, Middle East, Reliability, Facility Values, Discrimination Values, Post-test analysis, Reverse engineering, Learning outcomes

1. Introduction

As with other fields of educational assessment, the primary purpose of language testing, is to ensure ways of making decisions about testees regarded as fair and reliable by users of the assessment results. Within an educational context, test use within tertiary education is often employed as a gate-keeping tool, for countries wanting to ensure as many testees as possible have the opportunity to pursue their studies in higher education should they demonstrate the required ability. With any high-stakes assessment, it is considered essential to investigate the usefulness and usability of the test taken before going live. This ensures the test both genuinely elicits information regarding the construct of interest, and that any administration of the assessments provides consistent results (Fulchner 2010). Literature suggests this is achieved through field testing items, and subsequently critically evaluating specifications and items in light of pilot results (Bachman 2004; Fulcher and Davidson 2007).This all takes place before an assessment goes live. Such practices will be standard

for well-funded international and national exams, but perhaps less so for individual higher education institutions. This was the situation EAP instructors faced at UKH. The combination of limited time, and a body of students to use who did not know the test population proved an obstacle.

This paper describes the implementation of post-test analysis to maintain reliability and confidence in scores. I argue that simple steps can be taken to assure quality assessment. The paper concludes with core recommendations that must be in place for teachers and institutions in Kurdistan to achieve greater assessment confidence.

2. Background

In the 2012/2013 academic year, the University of Kurdistan-Hewler provided a two-year foundation program, comprised of English language development and core academic skills. The assessment format for this program was summative assessments at the end of semester per module, with students sitting a final English Exit Exam at the end of the two years.

Following the end of the academic year, the foundation program underwent review and development. The impetus for this was a university-wide review initiative, but the decision was motivated by the year's academic results. Of the 2012/2013 foundation cohort, 72 students were terminated and a similar number transferred to other higher education providers in the KRG. Some students rejected their score and decisions about their progression, and student demonstrations were reported in local media in Kurdistan.

It became clear there was a perceived lack of confidence in the assessment process by some students. Part of the universities response to media coverage was to initiate a transparency committee. This in turn inspired the author to look at how this initiative could be promoted in the assessment practices of the English Foundation program.

3. The Reading Assessment

The 2013-2014 program now consists of three cycles of assessment for learning during the semester, plus the traditional end of semester assessment of learning of each module: Writing, Reading and Listening. The three cyclical assessments during the semester, each of which were four weeks in length, help inform students of their progress and to enable them to identify gaps in their knowledge. The Reading module has been used to exemplify the implementation of post-test analysis. The Reading exam cycles are internally mandated achievement tests, in that content of the assessment is associated with instruction, looks to the past and is a measure of what has been learned (Fulchner 2010; MacNamara 2000:6). The module assessments consist of one vocabulary and one comprehension exam. This paper will examine vocabulary cycle assessments for cycle 1. The level tested is B1 on the CEFR, with scores of 50% and above being determined as equivalent to this. This posed the first problem associated with interpreting scores, and thus confidence in them.

The overall pass mark of 50% proposed at UKH is roughly equivalent to an IELTS 5, or B1 on the

CEFR scale. The test provided to the students, and the level of the module was pitched at a B1 level. This would mean 50% pass mark of a B1 test, would be roughly be A2 on the CEFR. Thus, the higher the score on the reading cycle test, the closer to B1 we could generalize students to be. In effect, students would need to score above 80 or 90% to be labelled as having attained B1. This issue needed to be addressed.

Cycles of instruction were theme-based to contextualize the most frequent 540 items on the Academic Word List (Coxhead 2000). The core aim of the module was to increase students' vocabulary threshold. The specifications for the assessments were very light, something the author recognizes as an area in need of further development. The more detailed the test specifications are, the greater the chance of greater validity and reliability. The assessments followed the same theme as the cycle and assessed each item taught in that cycle; typically this was 40-50 items. Items were designed to test receptive knowledge. Students were given 1 hour to complete the test.

4. Method

The post-test analysis was achieved using a critical reverse engineering approach (Davidson and Lynch 2007). Here, item performance is reviewed in light of the test purpose and analyses of item-spec congruence, as well as how useful an item is in relation to other items on the test. It must be noted when such analysis is applied to make assessment revisions, it is traditionally done before the next application of the adaptedtest. However, the author used this analysis to identify poorly performing items or sections that significantly affected score reliability. The author's aim was to ensure only students at a B1 level of attainment would pass and to prevent weaker students from passing due to poor test design. Poor items were removed from the scoring of the assessment, and new grade totals were calculated for students. Thus the test provided raw scores, which were then modified after post-test analysis. It is this method which the author advises practioners to apply in similar contexts.

Overall test performance was reviewed through exploring i) central tendency, ii) content validity, and iii) item analysis. Item analysis involved looking at Facility Index (F.I.) and Discrimination Index (D.I.), and using the Cronbach's Alpha reliability coefficient to measure internal consistency. Outliers with scores of zero were removed to prevent skewing of data.

When making decisions on item statistical characteristics elicited for Facility and Discrimination values, test specifications should ideally pre-state acceptable characteristics. This was not in place at the time of running the first vocabulary test in cycle 1. Literature suggests, acceptable F.I.s range from 0.3 to 0.7 (Henning 1987:50 in Fulchner 2010). However, for achievement tests, we could only consider values of between 0.65 and 0.75 to be acceptable. Anything greater than 0.8 will not provide enough spread to discriminate testees, and anything below 0.6 provides more spread than we would expect for an achievement test, given that content reflects what has been taught.

For D.I., literature suggests items writers should be satisfied with indices of above 0.3(Buck 2004:132). As noted earlier for an achievement test, we want to discriminate between abilities, but

only for the purpose of identifying who has reached the accepted level of competence. The author suggests indicies valued from 0.3 to 0.45 indices provide acceptable discrimination for a negatively skewed curve of the sort expected for an achievement test. Finally, acceptable Cronbach's Alpha reliability coefficient is ranged from 0.9 to 0.99 for vocabulary tests according to Lado (1961 in Hughes, 2003:32).

5. Results

This section will describe results and the recommendations that were made to test 1. This demonstrates how the author tried to maintain confidence in scores, but also implement the important process of test improvement.

Statistics

TotalScore

N	Valid	89
	Missing	0
Mean		26.2022
Median		26.0000
Mode		25.00
Std. Deviation		6.89727
Skewness		-.184
Std. Error of Skewness		.255
Kurtosis		-.759
Std. Error of Kurtosis		.506

Figure 1. Cycle 1 Vocabulary Test

For Test 1, descriptive statistics for overall tests scores are given in Figure 1. Examination of this shows a mean score of 26. Given that the pass mark was 50%, 75% of the students passed. Thus it appears students are doing well after the first cycle of the semester. However, this is a surface level inference. It is important to look more closely at other statistics. Was the test too easy? Is a pass mark of 50% perhaps too low, as discussed earlier?

In order to make this judgment, descriptive statistics of central tendency can be an initial indicator. The Standard Deviation was a little high at 7, with a ratio of 27%. This suggests the test was not easy and produced a bigger spread of students than we would like for an achievement test. In fact, it resembles a proficiency test. This is confirmed in the in curve direction and height. When viewing the histogram in Figure 2, it does not visually resemble that of a significantly negatively skewed curve as we would expect from an achievement test. In fact, it resembles a proficiency test. We see a skewness of -.184., which though negative is not significant enough to say that the test was far too easy. Nevertheless, the curve is not mesokurtic, confirmed by the kurtosis of -.759 in Figure 1. While a negative value usually implies a leptokurtic curve, or discriminates well for an achievement test, the absolute figure is not significant as is less than one Standard Error of Kurtosis (Brown 1997:20). In summary then, based on overall score statistics, it does not appear the test overall has been too easy, but it does not indicate the performance expected of an achievement test. Thus, while students have met the learning outcome of 50%, it appears the test population have not really

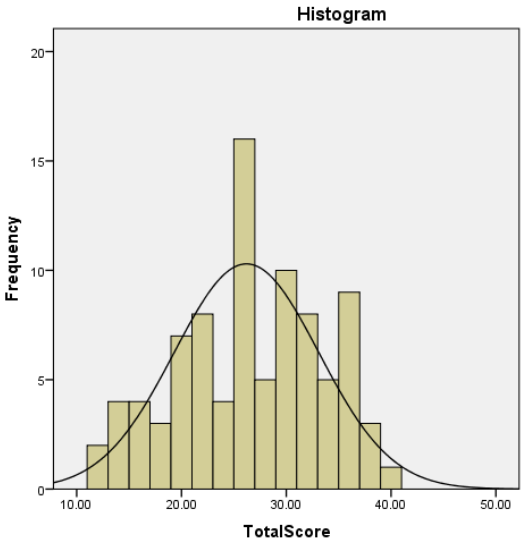

Figure 2. Cycle 1 Vocabulary Test

acquired much of the vocabulary studied, especially considering the discussion above of what the pass mark of 50% really means.

More useful indicators of test difficulty are found in F.I. and item discrimination (See Appendix 1a and 1b). The mean facility value for the test was 0.62, suggesting the test was slightly easier than acceptable ranges of 0.65 – 0.75, but not significantly. The mean D.I. was 0.3, just on the border of what is acceptable. The next question to ask is whether a section or certain items were perhaps too easy, as mean values can hide issues. For F.I., the mean for section 1 and 2 were around 0.75, at the top end of acceptable limits, but within each of these sections, half of the questions were in 0.8 to 0.9 ranges, thus too easy. This is evident in the D.I. as both sections scored around 0.2, thus weaker students were doing just as well as top level students on the majority of the questions, something we would not expect when considering overall test scores. Section 3, however, was much different. The mean F.I. was 0.48, meaning the students found this section particularly challenging. The D.I., though was much better on this section with the average being 0.4. Thus it is this section that really demonstrated which candidates should pass, and those who should not. Test 1, then, did not discriminate well in sections 1 and 2.

In terms of content validity, Part 1 (See Appendix 2a) was an item discrimination task where students were required to choose from 3 given options the best synonym for a given word. Students were provided a sentence that put the word in context for them. For Part 2 (See Appendix 2b), students were required to provide a synonym for a given word, without context provided. For Part 3 (See Appendix 2c), students were required to complete a gap with an appropriate word from a selection provided. There were multiple words from different word classes, meaning the questions could not be completed with syntactical knowledge, but required semantic knowledge. What this means for overall content validity is that the same construct of synonymy was tested twice in sections 1 and 2. In discussion with teachers after the test, it was felt that the skills needed when using a dictionary to select the appropriate word for a given context had not been tested. This was something students were required to do during classes. It is fair to say that the content validity needs improvement, as not enough of what constitutes knowledge of a word had been tested.

Reliability Statistics

Cronbach's Alpha	Part 1	Value	.558
		N of Items	22[a]
	Part 2	Value	.818
		N of Items	21[b]
	Total N of Items		43
Correlation Between Forms			.593
Spearman-Brown Coefficient	Equal Length		.744
	Unequal Length		.744
Guttman Split-Half Coefficient			.718

Figure 3.shows the Split Half reliability statistics

The overall reliability of the assessment was 0.82. This is lower than the minimum required for a vocabulary test of 0.9. The post-test analysis recommendations for test 1

with regards to scores were sections 1 and 2 be removed. This was supported by a split-half reliability analysis, seen in Figure 3, confirming it was the later part of the test which was more reliable.

This now meant the test was scored out of 20. This change saw only 47% of students attain a 50%. It should be remembered that the author questions the value of this score, in that as a pass mark it is too low, and that from overall student performance, we can see that students do not appear to have retained much of the vocabulary studied beyond being able to apply it to synonymy. However, we can now be much more confident in the scores of students from this test.

As for test revision, for the next cycle two new question types were designed. The first new task required 3 constructs (see appendix 3a). The item tests primarily sematic knowledge of the word, and also a small element of syntactic knowledge to identify the word class. The third construct was deducing meaning from context, as a number of words with similar meanings may be possible, but not appropriate for the context provided.

It is perhaps this last construct which students struggled with. A lot of support is provided for students in learning word meaning, and the other two parts of the exam, although not reliable, seem to indicate that students have rote learned synonyms, but not the actual semantic value of the word contextually. This is a higher order knowledge of vocabulary. Thus, the second new item type was designed to assess students' ability to match appropriate meaning to context (see appendix 3b).

The resolution to the 50% pass mark issue raised earlier was resolved through multiple cut scores and conversion. Four scoring bands were created: *Above Standard, Standard, Below Standard A, and Below Standard B. Standard* meant that students were on track to achieve B1 by the end of the module. The cut scoreswere set at 25% intervals, thus the pass mark was removed, in place of the symbol system above. This was intended to focus student attention away from a score focus that ignored feedback and remedial planning, to consideration of what they needed to do next in their learning journey. If students scored 75% or above they were considered as on target. Students scoring below were divided into two groups: *Non-Standard A* who needed optional remedial support, and *Non-Standard B* was provided mandatory support. However, as it was communicated to students that the pass mark was 50% and the banding put in place during the first cycle, a score conversion was needed. If a student scored 75% or *Standard,* this was converted to what 50% was on the global scale of the CEFR discussed earlier. Through the more appropriate cut score for Standard, only 22% of students were in fact on track and doing well, not 75%, or even 47% after post-test analysis.

Following the analysis at the end of cycle 1, a report for the academic board detailing results and recommendations was produced and shared. The same analysis was conducted for the Writing module with regards to inter-rater reliability. The Vocabulary assessment report was intended to be shared with the student body with accessible rationale for why certain decisions about scores had been made. The intention was to promote transparency with all stakeholders. Sadly, this was never shared with teachers, nor students. Such analysis was also not conducted for other testing cycles

during the semester, and no such statistical quality assurance is in place for the second semester in any module on the Foundation program.

6. Conclusion

This paper has demonstrated that post-test analysis of statistical characteristics of items and sections is essential to ensure that scores are both meaningful and enable decisions makers to actwith confidence. Regardless of the 50% pass mark issue, had no analysis been conducted, the 75% of students would have mistakenly been labeled as successful, when in fact only 22% of students were on track to pass.This paper has also shown that by generating such data, classroom-based feedback can be produced which shapes future lessons and materials. The two new assessment task types were introduced in class, and better supported students with the learning and skills of context negotiation and meaning.

The author accepts though, that there are indeed issues with this approach that need exploring outside of this paper. There are ethical considerations of revising grades post-test based on data highlighting issues realted poor test design. This is clearly not the fault of the students. However, it is also unethical not to quality assure tests and respond to issues when identified. This paper did not explore other potential issues that may have resulted in certain item performance. Poor discrimination can be the result of cheating, or good teaching of weaker students. This could have been be explored through pre and post-test surveys of students and teachers.

The recommendations for colleagues in the KRG are simple. Analyse central tendency inrelation to test purpose, and calculate Facility, Discrimination and reliability values at the very least after each live test you conduct.

References

Bachman, L.F. (2004) *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Fluchner, G. (2010). *Practical Language Testing*. London: Hodder Education.

Fulchner, G. & Davidson, F. (2007). *Language Testing and Assessment: An Advanced ResourceBook*. Oxon: Routledge.

MacNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Brown, J.D. (1997).Skewness and kurtosis.*Shiken: JALT Testing & Evaluation SIG Newsletter*Vol. 1 No. 1 Jan 1997 (p. 20- 23).

Item Statistics

	Mean	Std. Deviation	N
Q1	.708	.4573	89
Q2	.854	.3552	89
Q3	.652	.4791	89
Q4	.865	.3435	89
Q5	.809	.3953	89
Q6	.820	.3862	89
Q7	.697	.4623	89
Q8	.663	.4754	89
Q9	.494	.5028	89
Q10	.652	.4791	89
Q11	.775	.4198	89
Q12	.910	.2876	89
Q13	.831	.3765	89
Q14	.697	.4623	89
Q15	.820	.3862	89
Q16	.719	.4520	89
Q17	.798	1.1888	89
Q18	.607	.4912	89
Q19	.764	.4270	89
Q20	.191	.3953	89
Q21	.899	.3032	89
Q22	.719	.4520	89
Q23	.798	.4040	89
Q24	.404	.4936	89
Q25	.528	.5020	89
Q26	.438	.4990	89
Q27	.416	.4956	89
Q28	.337	.4754	89
Q29	.528	.5020	89
Q30	.382	.4886	89
Q31	.719	.4520	89
Q32	.787	.4121	89
Q33	.438	.4990	89
Q34	.360	.4826	89
Q35	.326	.4713	89
Q36	.404	.4936	89
Q37	.337	.4754	89
Q38	.685	.4670	89
Q39	.382	.4886	89
Q40	.697	.4623	89
Q41	.596	.4936	89
Q42	.697	.4623	89
Q43	.472	.5020	89

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q1	25.966	49.215	.263	.821
Q2	25.820	49.876	.220	.822
Q3	26.022	49.818	.157	.824
Q4	25.809	50.861	.026	.826
Q5	25.865	49.913	.186	.823
Q6	25.854	49.853	.203	.823
Q7	25.978	48.954	.300	.820
Q8	26.011	48.739	.323	.820
Q9	26.180	48.672	.311	.820
Q10	26.022	49.931	.140	.825
Q11	25.899	50.092	.141	.824
Q12	25.764	49.887	.280	.821
Q13	25.843	49.248	.325	.820
Q14	25.978	49.590	.200	.823
Q15	25.854	48.467	.464	.817
Q16	25.955	49.362	.243	.822
Q17	25.876	48.428	.077	.843
Q18	26.067	49.927	.135	.825
Q19	25.910	49.560	.227	.822
Q20	26.483	50.389	.100	.825
Q21	25.775	49.585	.335	.820
Q22	25.955	48.725	.345	.819
Q23	25.876	49.496	.255	.822
Q24	26.270	49.449	.204	.823
Q25	26.146	48.353	.359	.819
Q26	26.236	47.682	.461	.816
Q27	26.258	47.671	.466	.816
Q28	26.337	48.681	.332	.819
Q29	26.146	47.694	.456	.816
Q30	26.292	49.959	.132	.825
Q31	25.955	49.203	.268	.821
Q32	25.888	48.851	.362	.819
Q33	26.236	47.773	.447	.816
Q34	26.315	48.423	.365	.818
Q35	26.348	50.343	.081	.826
Q36	26.270	48.017	.416	.817
Q37	26.337	48.271	.395	.818
Q38	25.989	47.579	.514	.815
Q39	26.292	47.686	.472	.815
Q40	25.978	47.818	.481	.815
Q41	26.079	47.937	.428	.817
Q42	25.978	47.863	.474	.816
Q43	26.202	47.572	.474	.815

Appendix 1b. Discrimination Values shown in column 4.

Appendix 1a. Facility Values shown in column 2.

Appendices 2

Appendix 2a – Example task from section 1 of Cycle 1 Vocabulary test

Section One **(18 points)**

Read and circle the most appropriate word to complete the following sentences.

Example: Eating a diet high in saturated fats (confirms / consumes / poses) major health problems.

Appendix 2b – Example task from section 1 of Cycle 1 Vocabulary test

Section Two *(20 points)*

Write a more formal synonym for these informal words, using the AWL words from task cycle 1.

Example: separate ____discrete_____

Appendix 2c – Example task from section 3 of Cycle 1 Vocabulary test

Section Three (20 points)

Choose the appropriate word from the box to complete the sentences.

Example: It is not good to ___rely___ on carbohydrates only as a source of food.

pose rely require need

Appendices 3

Appendix 3a – Example task for section 1 of Cycle 2 Vocabulary test

Read the definition. Circle the word that is defined in the sentence provided.

Example:

0. The land or buildings of a university or college where the students study.

"Look around the sprawling Chaparral High School campus in Scottsdale, Arizona at lunch time, and the social geography of the 1,850 students is clearly instituted."

Appendix 3a – Example task for section 2 of Cycle 2 Vocabulary test

Read the sentence provided. Circle the number of the best definition for the underlined word

0. *Adolescence has always been a time of* ***identity****formation, with inclusion and exclusion, trying out new ideas, styles, and friends.*

1. somebody's name or who they are

2. the qualities or attitudes that a person or group have, which make them different from others

3. exact similarity between two things