SHIKEN

Volume 18 • Number 1 • August 2014

Contents

Testing and Evaluation SIG

Foreword

Trevor A. Holster, J. W. Lake
TEVAL SIG Acting Publication Chairs, *Shiken* Acting Editors

As readers may already know, the previous *Shiken* editorial team found themselves overcommitted with other projects and were not able to continue beyond the publication of *Shiken Research Bulletin 17,* leaving our SIG without a permanent Publications Chair and *Shiken* editor. Therefore the TEVAL SIG executive committee appointed acting editors until a permanent editor can be found. One of the legal requirements of a JALT SIG is to publish a newsletter three times per year, reporting on SIG activities and other issues of interest to members. We have not met this requirement for several years now, primarily because of a lack of submissions to *Shiken*. It is our hope that this issue of *Shiken* will mark a reversal of this and that the contributions to this issue will illustrate the important place of testing and evaluation in language education.

In addition to losing our *Shiken* editorial team, two executive members with many years of service, Jeffrey Hubbell and Ed Schaefer, are also retiring this year, leaving us in desperate need of officers. The TEVAL SIG has a steady membership, indicating that many JALT members consider issues of testing and assessment important enough to pay membership dues, but we are struggling to attract more active participation. This is becoming a serious matter because if we cannot maintain the minimum number of officers required under the JALT constitution, we face the possibility of the TEVAL SIG being disbanded by JALT for not meeting legally binding obligations.

Talking with Jeffrey Hubbell, Ed Schaefer and other veterans of both TEVAL and the Japan Language Testing Association (JLTA) raised the point that a major concern when TEVAL was established in 1997 was the harmful effects of high-stakes tests, notably university entrance exams. Test misuse is still a major concern, but TEVAL has important positive contributions to make beyond raising issues of the negative consequences of test misuse. As the name suggests, JALT is an organization concerned with language education. Although many of our members conduct research, it is rare to meet a JALT member who is not primarily a language teacher, with research a secondary focus. Most teachers are required to assess their students and assign grades that often determine eligibility to graduate. However, many language teachers have little or no formal training in language assessment, so one of the key contributions that *Shiken* can make is helping teachers improve the quality of their classroom assessment, an area where much of the psychometric research on large-scale tests is not directly applicable. In many classroom situations, the psychometric properties of tests may be subservient to pedagogical issues such as the effects of high-stake tests on motivation, integration of classroom tasks and content into assessment, and the formative effects of assessment. In some ways, this makes classroom assessment more technically challenging than large scale proficiency tests, where the sole purpose is to measure a well-defined trait of ability as quickly and efficiently as possible. Unfortunately, classroom assessments generally do not meet the standards of replicability and generalizability demanded by research journals, so issues of classroom assessment are underrepresented in the literature.

All of the contributors to this issue of *Shiken* addressed issues of concern for classroom assessment. Trace and Janssen provide a tentative investigation of the relationship between item difficulty on two variations of cloze tests and measures of word association derived from corpus analysis, finding support for the view that the contextual information provided by analysis of word associations compared with simple word frequency can improve the selection of words for deletion. Crucially, the techniques used by Trace and Janssen are accessible to teachers with moderate technical ability who need to develop integrated assessment and instructional materials. Marshall documents the development of an assessment rubric for classroom use, illustrating the many difficulties inherent in the integration of criterion referenced

assessment in the classroom, especially of the difficulty of providing meaningful formative feedback to students. Cubilo provides an accessible discussion of the frequently misunderstood concept of validity and the nature of validity arguments. Given that language tests can determine entry into a school or program, or whether students are eligible to graduate, the implication that a validity argument presented for a test in one context may be irrelevant in another highlights our responsibility to gather evidence and present an argument justifying the interpretation and use of assessments beyond the citing of reliability coefficients or descriptions of test content. Dunkley interviews Dr Meg Malone of the Center for Applied Linguists (CAL) whose experiences with assessment literacy for language teachers highlighted differences in the information that language testers and language teachers considered important, leading to improvements to CAL's directory of language tests to make relevant information more accessible to teachers and administrators. Brown, continuing his *Statistics Corner* contributions that date back to the very first issue of *Shiken*, discusses the differences between norm-referenced and criterion-referenced tests from the perspective of classroom assessment, providing a practical introductory guide for teachers wishing to develop classroom assessments.

Finally, we are very happy to introduce *My Tests*, a new column developed at the initiative of Jeffrey Durand, the TEVAL treasurer and long serving member of the *Shiken* editorial board. The first installment of *My Tests* deals with the problem of designing a judging plan in rated performance tests, an issue that should be of concern for anyone trying to implement such tests on a program-wide basis. *My Tests* is a forum for our members to share experiences with tests and to seek and provide advice. In test development we often encounter problems that must have been encountered before, but are forced to reinvent solutions because there is no forum to share our experiences. This forum is thus intended to help our members, both through the answers that you share and in the questions that you ask.

Thank you to all our contributors and our readers, we hope you find this issue of *Shiken* valuable and look forward to seeing you all at the TEVAL booth in the SIG area at the upcoming JALT International Conference in November.

Trevor Holster and J. Lake

Acting *Shiken* Editors

Corpus-informed test development: Making it about more than word frequency

Jonathan W. Trace[1] and Gerriet Janssen[1, 2]
jtrace@hawaii.edu
gjanssen@hawaii.edu

1. University of Hawai'i at Mānoa
2. Universidad de los Andes–Colombia

Abstract

Given the rising popularity and usefulness of corpora in the field of applied linguistics, more and more there is a need to identify practical applications of the different tools available beyond just word frequency. One area where corpora seem ideal for this is in the realm of second language assessment. This study looks at the use of corpus-informed test items on an academic English vocabulary test (N = 203). Two different formats of the test (c-test and multiple-choice) are analyzed to explore possible relationships between item characteristics for difficulty and contextual information. First, Rasch measurement is used to determine the difficulty of a set of common items across both tests. These results are then compared with a series of mutual information scores based on collocations and multi-word constructions with the target items. The goal is to examine possible relationships between context and item difficulty, and more importantly provide teachers and test-designers with one way to utilize corpus linguistics to create more effective language assessment tools.

Keywords: corpus linguistics, language testing, formulaic language, vocabulary

Introduction

Corpus-based research is still a growing area in applied linguistics, though it boasts a long and productive almost 40 year history, with studies ranging from basic descriptions of language (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999; Sinclair, 1990), to more practical applications such as lexico-grammatical approaches to instruction (Liu & Jiang, 2009) and materials development (Chang & Kuo, 2011). One area that is slowly building momentum is the use of corpora in language assessment design and use (see Barker, 2005; Coniam, 1997; Sharpling, 2010). These two areas would seem to go hand in hand with one another given that both espouse such concepts as reliability and authenticity. Yet it still seems that most of the testing literature that incorporates corpus is still limited to conversations that seldom go beyond word frequency (Crossley, Salsbury, McNamara, & Jarvis, 2010).

Certainly there is more value to be had from these vast databases of authentic language than just how often a word is used? As teachers and test designers, we know there is more to knowing a language than just the individual parts on their own. Rather, it is language in use, with considerations of context or lexico-grammatical function that we should be interested in measuring and teaching. Even when our focus is on something narrow, such as the vocabulary test in this study, there are still several distinct constructs that stand out as important for successful mastery of a language that go beyond frequency, such as vocabulary depth (Nation & Snowling, 1997). Neither are we always interested in knowledge of words on their own, as language is not given to us in piecemeal but as part of a larger whole. This includes knowledge of formulaic language (e.g., *n*-grams or idiomatic expressions), which have been common topics of discussion in the field (Evert, 2009; O'Keeffe, McCarthy, & Carter, 2007) but remain relatively unexplored in the field of language assessment.

The goal of this study is to highlight one possible way of expanding our use of corpora in the field of language assessment from a very practical and authentic position. Using a vocabulary test, this study will explore one method of analyzing words in context, as well the outcomes of a corpus-informed test in the

hopes of providing teachers and test designers tools to better measure language. To this end, the following research questions were asked:

1. What if any are the relationships between item difficulty and mutual information as identified by corpus-derived data?
2. To what degree are these relationships similar across different test formats?

Methods

Participants

Data were collected from responses on a vocabulary test from a total of 203 examinees at a South American University in Colombia. The test is part of a larger placement exam for incoming Ph.D. students in an English for Academic Purposes support program. Examinees were primarily L1 speakers of Spanish and ranged from low beginners to advanced users of English.

Instruments

The assessment tool discussed in this study was a test of academic English vocabulary. Along with writing and speaking subtests, the vocabulary test was part of a new reading pilot that included sections for grammar and reading comprehension. Data for this study were collected from three administrations of the vocabulary subtest.

As this test is still in a piloting stage, revisions to the test are ongoing and variations exist across each of the three administrations examined here. Most apparent among these is that the first two administrations (n = 124) used a c-test (CT) format, where examinees were given a passage with several missing words that they are required to supply. Unlike a traditional cloze procedure, in a CT the first letter of each missing word is provided both as a clue for test takers, as well as to limit the possible number of accepted responses. For the final administration (n = 79), the test was converted to a multiple-choice format (MCT) based on apparent difficulty problems with earlier versions of the test.

The original design of the test incorporated a 500-word passage with 26 missing words using a rational pattern of deletion. Items were selected based on corpus data from the *Corpus of Contemporary American English* (COCA; Davies, 2008) using statistics such as word frequency and mutual information scores. As the researchers wanted to choose a topic that was academic but also general, to avoid biasing any particular academic background or major, an article on the history of the wheel was selected from an online academic journal.

After the first administration of the test, item analysis was conducted using classical test theory. Six items were removed from the test as problematic and two new items were introduced (k = 22). Results of the revised test were also analyzed, and based on these findings the researchers decided to change to a MCT format. Examinees were presented with four possible choices in-text and were required to circle the correct answer rather than producing any language. The new test removed ten of the previous items and added eight new items (k = 19). In total, twelve items were shared across all three tests, and these were used as the basis for the final linguistic analysis detailed below.

Procedure

Because of our interest in the relationships between an item's target word and the immediate context surrounding that word, mutual information (MI) scores were used as reported by the COCA. MI is a statistical measure of association that indicates the degree to which a set of words or a phrase is likely to

appear in the same pattern together in the language (Biber, 2009; Evert, 2009). It works by comparing the frequency of a multi-word pairing or phrase. A high MI value is found when there is a strong likelihood of words or a phrase to appear together within the corpus. According to Davies (2008), MI values of 3.00 or higher indicate a high chance of a set of words being bound together semantically in naturally occurring language. Scores are dependent upon the individual word frequencies, as those words with very high frequencies show up in many different contexts and their appearance with other words may be more due to random chance than any kind of formulaic-ness.

In order to measure collocations between the target word and nearby function words, MI scores were gathered for all function words within a fixed area around the target word. While MI is typically used to look at fixed semantic phrases (e.g., *strong coffee*), taking into account only immediate pairings of words, it seems logical that words that are still local (e.g., within the same *T*-unit) but not immediate might also have a triggering effect for a particular target word, and this has been explored in the psycholinguistics literature (e.g., Duffy, Henderson, & Morris, 1989). To reflect this, analysis of MI scores included all function words within four collocations to the left and right of the target word within the same *T*-unit. In order to make comparisons about the relationship of collocation on item difficulty, the maximum left and right MI scores were used in the analysis.

Biber (2009) differentiates between collocations of function words and multi-word formulaic sequences. Given our interest in how different kinds of context influences item difficulty, it is important to consider MI values for both of these kinds of constructions. As with collocations, multi-word formulaic sequences required a bit a preparation to measure in a systematic and authentic way. Formulaic sequences can be quite varied, both in relation to the content words within a fixed set of function words (Renouf & Sinclair, 1991), as well as in length (e.g., *fairly certain* vs. *fairly certain that*). As the target items were typically function words, and both the CT and MCT formats limited the number of possible answers, the main concern was determining what is or isn't part of a phrase. One possible way of accomplishing this is to look at multi-word constructions in three areas: (a) before the target; (b) after the target; and (c) including the target. By isolating these three patterns, we can check MI values for constructions with the target word as the base, and then expand outwards in the appropriate direction. Different number *n*-grams can be compared in terms of their MI scores, and the construction with the highest MI score and fewest number of words can be reasonably identified as a multi-word formulaic pattern in the data.

Analysis

Exact answer scoring was used in the analysis of the item data as a way of controlling for differences in responses by examinees. While more than one answer was possible for some of the items, exact answer scoring allowed us to focus only on the original constructions in their relationship to item difficulty and corpus linguistic features. Given the different formats and modes (e.g., receptive vs. productive) of the CT and MCT, analyses were conducted separately. For the CT analysis, only those items that were shared across both tests were included ($k = 20$). As there was only one version of the MCT, all 19 items were included in the analysis.

Rasch measurement was utilized to analyze item responses on both tests using *Winsteps* (Linacre, 2010). Unlike classical test theory, Rasch, which belongs to the item-response theory family of test analysis, can give a sample-free estimation of item difficulty as it relates to examinee ability levels along a true interval scale. The benefit of Rasch modeling has been discussed extensively in the literature (e.g., Henning, 1984; McNamara & Knoch, 2012), but suffice to say this method of analysis provides a more generalizable and readily comparable interpretation of item difficulty.

Corpus analysis of the 12 common items on the test was carried out according to the procedure outlined above, with five sets of MI scores for each item: (a) left MI; (b) right MI; (c) pre *n*-gram MI; (d) post *n*-

gram MI; and (e) mid *n*-gram MI. In addition, the frequency of the target word per one million words was recorded. While the COCA provides both spoken and written data, as these were items on a vocabulary test and measuring knowledge of the written language, only written corpus data was used for all analyses.

Results

Descriptive data for both test formats are displayed in Table 1, including means, standard deviations, minimum and maximum values for each the CT results and the MCT results. Notice that the mean score on the CT was markedly low (M = 3.90) with a relatively low degree of variation in scores (SD = 3.70), indicating that the test was quite difficult for the sample of examinees. By comparison, the MCT was more normally distributed (M = 12.90, SD = 3.86), with examinees appearing to perform much higher than on the CT. Cronbach's alpha reliability estimates are also included in the bottom row of Table 1, indicating the degree to which the scores on the test were internally consistent. Scores on the CT were more reliable, with an estimate of 88%, while the MCT was slightly less reliable at 73%. We should be careful in over-interpreting the reliability of the CT given the low degree of variance of scores and positively skewed distribution. These might be causing this value to be higher than it actually is, as reliability estimates work under the assumption of normally distributed data. A lack of variance might mean that the scores are consistent, but only consistently low, and have little to do with actual test function.

Table 1
Descriptive Statistics for the Vocabulary C-Test and Multiple-Choice Test

	CT	MCT
M	3.90	12.90
SD	3.70	3.86
Min	0.00	3.00
Max	14.00	19.00
N	124	79
k	20	19
α	.88	.73

Item analysis of the tests was performed using Rasch analysis, which displays item difficulty as logit measures. Measures are spread across an interval scale with a mean value of 0.00, ranging negative (less difficult) to positive (more difficult). A first step to determining item function in Rasch is to check the fit of the items, or the degree to which the item measures can be adequately predicted by the model. Misfitting items were determined by evaluating infit mean square values and identifying items more than two standard deviations from the mean (McNamara, 1996). A preliminary analysis found that one of the common items misfit the model for the CT, and so was removed from the final analysis for a new total of 11 items.

Table 2 displays results for the 11 remaining common items in the order they appear on the tests. The first column displays the target word, followed by item difficulty in logits. Corpus statistics are also included for each item, including the frequency of the target word per one million words in the COCA, the highest MI value for near context words to the left and right of the target word, and *n*-grams with the target for left, right, and surrounding context.

Looking first at the item analysis data for the 11 common items, we can see that there are clear differences between difficulty measures for both tests. Notice that in general measures for the CT were higher than those for the MCT, which again points to the CT being the more difficult test. While the CT had four items with logit measures above 2.00, the most difficult item on the MCT was for the target word *people*

(1.96). Most items on the MCT were either closer to the center of the scale or quite easy as reflected by high negative values.

Table 2
Item and Corpus Statistics for 11 Vocabulary Items

Item	CT Measure	MCT Measure	Word Freq*	Left MI	Right MI	Pre *n*-gram MI	Post *n*-gram MI	Mid *n*-gram MI
Exactly	-1.88	-2.28	74.94	5.97	4.19	7.29	8.06	10.12
Certain	2.38	-0.87	130.01	6.61	2.17	5.47	0.00	5.88
Beneath	3.37	0.22	47.55	3.27	2.11	2.77	0.00	5.18
Learned	-0.86	-0.46	93.45	4.42	3.65	1.09	6.56	8.88
People	-0.06	1.96	1006.95	1.47	0.00	0.00	0.00	0.00
Before	-3.55	-2.05	668.49	4.13	3.94	3.33	6.46	8.21
Indicate	0.31	0.67	36.23	4.60	5.39	0.00	0.00	10.27
Making	-0.20	-0.87	227.06	1.70	3.36	3.76	7.09	0.00
Although	2.38	-0.19	239.89	0.00	1.70	0.00	0.00	0.00
Keeping	2.78	1.24	55.08	2.38	3.30	3.17	2.96	7.69
Everyday	0.62	-0.10	18.84	3.87	8.33	7.86	7.81	10.22

Note. * Word frequency based on the occurrence of the target word per 1 million words in the COCA.

For the corpus data, we can see that the frequency of the target words was quite varied, ranging from about 18-1000 occurrences per million words, though most items had values below 250. MI scores for left and right collocations showed that most items had at least one semantically related word in the near context. Recall that MI scores of 3.00 or higher indicate a semantic relationship, and only *although* (1.70) and *people* (1.47) were below this threshold. As the former is a connecting device and not likely to be directly contextually linked, and the latter was the most frequent word, these results could be expected. We find similar patterns in the data for multi-word MI scores. Overall, the highest multi-word MI scores occurred when the target word was centered in the phrase.

Pearson product correlations were used to gauge the degree of relation between corpus data and item difficulty. Table 3 displays these results arranged by item difficulty, frequency, collocation, and multi-word formulaic sequences. Given the number of comparisons, an *a priori* alpha of $p < .002$ was used in determining statistical significance. As we might expect given the low number of items in our sample, none of the correlations were determined to be statistically significant, though there were some interesting trends worth exploring in the data.

Table 3
Correlation Matrix for Item Difficulty and Corpus Statistics for 11 Vocabulary Items

	CT Measure	MCT Measure	Word Freq*	Left MI	Right MI	Pre *n*-gram MI	Post *n*-gram MI	Mid *n*-gram MI
CT Measure	1.00	.56	-.41	-.27	-.25	-.13	-.66	-.28
MCT Measure		1.00	.19	-.50	-.28	-.54	-.61	-.30

Note. * Word frequency based on the occurrence of the target word per 1 million words in the COCA. A Bonferroni adjusted a priori alpha value of $p < .002$ was set to account for the number of comparisons.

Notice that items in the CT displayed the strongest relationship with multi-word sequences following the target word ($r = -.66$). A negative value indicates that as item difficulty on the CT increased, the likelihood that the target word was the beginning of a fixed phrase decreased. The same was true for items in the MCT ($r = -.61$), but expanded also to *n*-grams that preceded the target item ($r = -.54$). MCT difficulty also appeared related to collocations occurring before the target word ($r = -.50$). This seems to show a

possible relationship between item difficulty and the presence of fixed word combinations or multi-word sequences, and that the pattern of this influence might change depending on the test format.

Discussion

Based on the proposed corpus-driven methodology described above, results seem to indicate tentatively that there is a relationship between item difficulty and the degree to which the item is connected to nearby context in the form of fixed expressions. Based on the correlation data, it appears that different contextual features are affecting difficulty as a whole across both tests. Difficulty seems to be influenced when the target is part of a multi-word phrase, either with a fixed sequence of words preceding or following the word. Alternatively, the degree to which single word collocations are related to item difficulty is not as clearly displayed, especially for the items in the CT. Word frequency was only weakly related to item difficulty, and only for the CT. This is likely due to the small sample of items in the analysis, but still worth nothing that collocations were more related than word frequency alone, as we might expect in a test of vocabulary in context (Crossley et al., 2010; Zareva, 2007).

Findings related to the CT and MCT also showed some apparent differences in formulaic language and item difficulty across test formats. As mentioned, the difficulty of the items on the CT seemed to be mostly unrelated to the presence or absence of collocations to the target. There was evidence of a possible relationship, however, between items on the MCT and the presence of collocations prior to the target. Items on both tests were sensitive to the presence of multi-word phrases, though again differences were found between test formats. While items on the CT seemed to be affected by sequences following the target, the MCT included sequences before and after the target. Neither was influenced when the target word was centered in a sequence, which was somewhat surprising as those values tended to be the highest and most common in written English (Biber, 2009).

It could be these differences were due in part to the presence of options in the MCT. Examinees might have been able to use the provided answer choices to help interpret the context, whereas in the CT examinees didn't have access to this added information and had to work from the context alone. We might think this would increase the effect of multi-word sequences on item difficulty on the CT, but the data doesn't seem to support this notion. This might be a result of the CT being too hard, or that examinees lacked knowledge about the context to make these kinds of judgments without more information. Unfortunately, without more information or better functioning items, it is impossible to be certain.

Conclusions

The goal of this study was to display one possible practical application of corpus-based test design through the use of different statistical procedures. While there remain a variety of questions about how corpus-informed tests function in different contexts, we hope that this can be a starting point for test designers to make more informed decisions when creating and selecting items.

This was a small-scale study with only a few items, and because of this we must be careful with the kinds of conclusions that can be drawn. That said, the results do seem to indicate possible benefits in using collocations to influence item difficulty, and point to the value in looking beyond frequency or individual words when testing vocabulary as authentic language use.

It is hoped that this information can lead to more in-depth studies of the use of corpora in test development. The next step in this research will be to look at a broader range of items that can more fully encompass different constructions of vocabulary in context, as well as incorporate eye-tracking methods to examine where test takers are looking when responding to items in a test, using online measures of processing to better understand the degree to which examinees use context in reading assessment.

References

Barker, F. (2005). *What insights can corpora bring to language testing? CRILE* (pp. 1–4). Lancaster University. Retrieved from http://www.ling.lancs.ac.uk/groups/crile/docs/crile%20lectures/barker0105.doc

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275–311. doi:10.1075/ijcl.14.3.08bib

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Chang, C.-F., & Kuo, C.-H. (2011). A corpus-based approach to online materials development for writing research articles. *English for Specific Purposes*, *30*(3), 222–234.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *Calico*, *14*(2), 15–34.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580.

Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition 15*, 791–801.

Evert, S. (2009). Corpora and collocations. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 1–53). Berlin: Mouton de Gruyter.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing*, *1*(2), 123–133.

Linacre, J. M. (2010). *A user's guide to Winsteps*. Chicago, IL: Author.

Liu, D., & Jiang, P. (2009). Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *Modern Language Journal*, *93*(1), 61–78.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. doi:10.1177/0265532211430367

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*, 359–370.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching* (pp. 1–333). Cambridge: Cambridge University Press.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128–143). London: Longman.

Sharpling, G. P. (2010). When BAWE meets WLT: The use of a corpus of student writing to develop items for a proficiency test in grammar and English usage. *Journal of Writing Research*, *2*(2), 179–195.

Sinclair, J. (1990). *Collins COBUILD English grammar*. London: Harper.

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research*, *23*(2), 123–153.

Diagnosing Students' Proficiency on a Spoken Performance Assessment

Paul Anthony Marshall
paulanthony.marshall@gmail.com

Abstract

The aim of this study was to diagnose specific gaps between current student proficiency and a target standard of proficiency in presenting a daily bulletin, in order to make an informed decision about what I can do to help students to span these gaps. After much trial and error with a variety of diagnostic tools, the study uses a thematic chart to successfully identify gaps in student ability on the performance assessment. Here, I begin by outlining the methodology, which has been broken down into three stages: definition of performance criteria, rubric development, and rubric operationalization. I then go on to reflect on the successes and shortcomings of the process and the decisions made.

Introduction

I was recently teaching English for Specific Purposes to eighty Laotian nationals at an Australian-managed gold and copper mine in southern Lao P.D.R. The management of the Training Department decided that the "Professional English" course should switch to using the Australian vocational performance assessment system of competency-based assessment which essentially meant assessing students on practical work-related tasks such as meetings or presentations. This decision was made part-way through the course which had already been fully planned and partly delivered so I needed to develop an effective approach, and quickly. Based on articles that I had read on formative assessment, I saw it as a possible vehicle to drive my students to success on performance assessments.

The first stage would be to build up a clear picture of my students' current levels of proficiency and of a realistic target level of proficiency. In order to do this, it would be necessary to carry out a formative assessment of students completing the task. As assessment criteria did not yet exist, I first set out to determine appropriate criteria. Ideally, these would be criterion-referenced in order to assess students according to an external, standardised set of criteria, which have been tried and tested.

Sadler (1989, p. 119) focused on "the nature and function of formative assessment in the development of expertise" where "student outcomes are appraised qualitatively using multiple criteria" and discusses the benefits and drawbacks of qualitative judgments, the use of descriptors, fuzzy, as opposed to sharp, criteria, and metacriteria, the criteria for using criteria. It provided guidance for many of the micro decisions made in this study. Black & Wiliam (1998) provided excellent procedural input for the implementation of formative assessment in the classroom which I used while planning the initial diagnostic stages. Huhta (2008) deals with the nuances between the definitions and functions of a variety of assessment types, and also introduces the idea of diagnostic competence which led me to use video to record student presentations. Biehler and Snowman (1997) contributed understanding of the importance of measurement and evaluation in the process of performance testing and during the analysis of test results. Davison and Leung (2009) supplied an insightful exploration of possibilities for using assessment for learning in the classroom.

While all of these articles provided inspiration and methodological input on utilising formative assessment to improve student competence on performance assessments, this study focuses only on the initial step; namely that of diagnosing areas of weakness for potential focus for formative assessment techniques. My research into the diagnostic evaluation of student presentations also consisted of

collecting assessment rubrics and an instructional article by Simkins (1999), both of which I utilised to select the most suitable assessment criteria, write descriptors, and design rubrics for the specific task of presenting a daily bulletin in my specific context.

Method

Participants

Before testing out the criteria, I needed to select a manageable set of performance assessments to try them on. I took a number of factors into consideration when choosing three students to represent the entire population of eighty Intermediate Professional English students. I had been teaching most of the members of this course for almost three years, and I was confident that the three students were representative of the entire Professional English population in terms of gender, the range of ages, backgrounds, professions, and the range of competence in English fluency, comprehension, and presentation skills. I felt that three students was a sufficient number for a small-scale study, and choosing an odd number avoided the possibility of split results. I sat with all three students and explained to them what I was asking of them.

Instrument Development

In order to formatively assess students' performance assessments comprehensively, I first had to select or create some appropriate criteria. The most effective method of assessment I had experience of was IELTS speaking and writing examinations which use a nine-band rubric. IELTS Examiners attend standardisation training in order to make sure they are all interpreting the criteria in the same way. However, by personally assessing my students against criteria, the results of the data generation would hinge on my own concept of the target standard, and were based largely on my own independent evaluations of student performances. Assessing student performance against multiple criteria and based on a target standard determined only by the teacher is by definition subjective, "the teacher must possess a concept of quality appropriate to the task and be able to judge the student's work in relation to that concept" (Sadler, 1989, p. 121).

The process of designing a suitable instrument consisted of a great deal of trial and error. Before experimenting with a group of existing oral presentation skills rubrics I had gathered to assess videos of my students' presentations (McCullen, 1997; NCTE/IRA, 2004; Swinton, 2012), I excluded irrelevant or inappropriate criteria from them such as those related to presentation slides. Where similar criteria existed on more than one of the original rubrics, I selected those that I judged to be most relevant to my students in their context. While I favoured the idea of an IELTS-style rubric with comprehensive descriptors, I decided to use universal descriptors, knowing that I would rewrite the rubric in a substantial way after this initial trial run.

This process resulted in Rubric A, shown in Figure 1, which combined the most suitable success criteria from a range of oral presentation skills rubrics. However, after viewing the three videoed presentations numerous times using Rubric A, I felt that the universal descriptors were unsuitable for the task, and the criteria needed reviewing. I went on to try out several more rubrics which had a variety of formats and some alternative, but similar criteria. I hand-wrote notes onto these rubrics about their strengths, weaknesses, and suitability in order to further refine the rubric. Following Simkins (1999), I limited the number of criteria to four because this forces the designer to prioritise which are the most important. I grouped together similar criteria, and incorporated criteria-specific descriptors for the groups to create Rubric B, shown in Figure 2. Again following Simkins (1999, p. 23), I created four levels of descriptor for each criterion because three levels does not provide sufficient discrimination but more than four leads to splitting hairs.

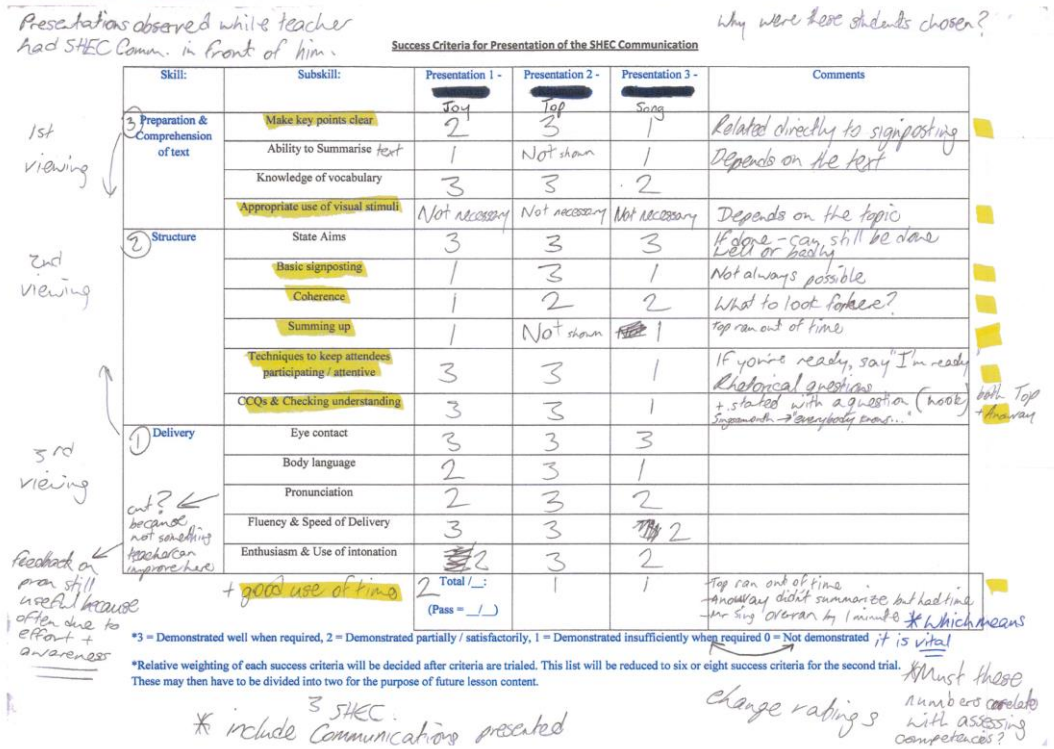

Figure 1. Rubric A

FINAL RUBRIC	1	2	3	4
Structure Coherence Basic signposting	Clear intro-content-summary / conclusion. Coherent and basic signposting used effectively.	Logical structure OR coherence / signposting.	Some structure OR insufficient but present coherence / signposting.	No identifiable structure. Incoherent. No signposting.
Getting Key points across Checking understanding	Made key points abundantly clear and made certain that they had been understood. Good awareness of audience understanding.	Made key points mostly clear, some checking of understanding OR one of the two points in box 1. (Awareness.) * Some awareness.....	Key points mostly unclear OR insufficient checking of understanding / awareness of audience understanding.	Key points unclear. No checking of understanding. No awareness of audience understanding.
Good use of time	Time used effectively for the purpose of content delivery.	Time used effectively but marginally ran out / overran / missed an opportunity.	Time used ineffectively – some content missed.	Important content missed.
Techniques to keep interest	A combination of rhetorical question, visual stimuli and ~~animated~~ personal style kept audience attentive.	Mostly kept audience attention through one or two techniques but could have been executed more effectively.	Kept audience interest some of the time but insufficiently.	Kept little or no audience interest.

(animated personal style = engaging style")

Not explains of main points
3 Little structure
3 coherent
3 No signposting
3 NO CCQs
3 No summary
3 basically reading with good intonation
no visuals fingers
1 signposting
2 clear not enough CCQs
2 No summary
as ✓ If you're ready, say "I'm ready!"
3
2/3 large void here
3 No summary
3
difficult and dull subject matter which was not chosen or organised by presenters.

Figure 2. Rubric B

Peer feedback

I discussed Rubric B with a colleague and received some brief feedback on it which can be seen hand-written onto it in Figure 2. I then used Rubric B to assess the three videoed presentations during repeated viewings, and hand-wrote very brief notes on student performance onto the rubric. This trial of Rubric B allowed me to evaluate the strengths and weaknesses of grouping criteria together, the descriptors I had written, and of my students' performances. I came to the conclusion that the grouping of criteria made assessment more difficult because frequently students would achieve one criterion but not the other in the same group. The descriptors did not allow for this eventuality. I also realised that limiting the number of criteria to four was completely unnecessary in this case because my purpose for the use of criteria was diagnostic and not to provide feedback or report progress.

Data collection

All of the evaluations were done by watching pre-recorded videos of student presentations. As a reaction to the results of trialling Rubric B, and after having started reading into data analysis and interpretation, I decided to alter the data generation process to evaluate the videoed presentations more thoroughly. While I was trying to decide how best to present my data, I considered presenting the comments in paragraphs by presenter, or in paragraphs by criterion but essentially I was searching for a method of presentation which, following Spencer, Ritchie, and O'Connor (2003, p. 210), allows searches to identify thematic categories and patterns and shown associations between phenomena within persons and between persons or groups of persons. As a result, I decided that it would be logical and easy to reference if this data could be searched by both presenter and criterion on one table, leading to the thematic chart shown in Table 1.

The thematic chart was not pre-planned; it was a contingency which I feel considerably improved the descriptive quality of the data gathered, which in turn facilitated my analysis of the data. The additional column for general comments about each presenter, and the additional row for general comments about each criteria meant that the data was not limited to my preconceived categories. I eventually prepared and processed my data and presented it in different formats to aid with analysis and interpretation, and to ensure it could be easily accessed and referred to.

I viewed the videos numerous more times while writing evaluative comments into the thematic chart for easy reference by criteria and by presenter. I was becoming very familiar with my students' presentations by this time which in itself meant that I could evaluate them in much more detail. I also included examples of actual presenter monologue where possible. Sub-dividing comments and monologue by specific criteria meant that I could specifically diagnose what students need training on, but it also served the additional purpose of categorising the data in preparation for analysis and interpretation.

On completion of the thematic chart, I assigned criteria to what I perceived to be the most enlightening four classes at a higher level of abstraction; questioning, emphasis, audience understanding, and time. Following this, I created an extra column at the end, and an extra row and at the bottom of the rubric. I used these to write a brief summary of the information included about each criterion, and about each presenter. This process aided both the analysis, and the interpretation of data.

One of the most useful and revelatory patterns that resulted from sorting and categorising my data was a possible insight into the thinking of the presenters. I discerned from the data that the presenters did not appear to assume responsibility for audience understanding. This can be implemented through asking questions to check understanding, emphasising key points, personalising, and concluding. The identification of this pattern will enable me to further observe this phenomenon, and to plan future lesson content based on this need.

Table 1

Thematic chart displaying assessment observations

Criteria: Students:	**1a Introduction: stating topic, activating schemata, creating interest**	**1b Rhetorical questions**	**1c Questioning the audience**	**1d Emphasis through repetition**	**1e Emphasis through stress**	**1f Emphasis through visual aids**	**1g Awareness of audience understanding and interest**	**1h Checking understanding of key points**	**1j Personalising / contextual-ising the content**	**1k ~~Summarising~~ Concluding**	**1m Good use of time**	**Comments on each presenter**
Joy	Joy stated the topic, then used a rhetorical question as a sort of hook to introduce the topic**. "So do you know how fires start from welding? OK, I can tell you now."** More would have been better.	Joy used one rhetorical question at the start**. "So do you know how fires start from welding? OK, I can tell you now."** More would have been better.	No other questions were asked. The opportunity to check understanding and / or contextualise the content was missed.	The key points were not repeated. This could have been an effective way of making sure the audience understood what the key points were.	Joy used intonation very effectively to keep audience interest, and to emphasise the key points.	No visual aids were used, but the content didn't necessitate the use of visual aids.	Very little awareness of audience understandings hown other than monitoring and maintaining interest by making eye contact.	This was not done despite finishing early. A missed opportunity.	This was not done. Joy could have asked the audience for personal experiences related to the topic.	This was not done. A missed opportunity to emphasise the key points through repetition, personalisation, or to check understandingof key points.	The missed opportunity to summarise or personalise the content (despite finishing early) was one of the main weaknesses of Joy's presentation.	Joy uses intonation, eye contact, and body language effectively but could benefit a great deal from using the other techniques listed here.
Top	Top introduced himself, stated the topic, & used a rhetorical question to spark interest. The question could have been more effective. He signposted **"today I'm going to talk about six ways..."**	Some rhetorical questioning. More would have been better. Top kept checking audience agreement with the points he was making by asking **"Yes?"**	This was done only briefly at the start**: "Do you think accidents are a kind of luck?" "Do you think that accidents can be prevented?"**	Top's checking of audience agreement was a method of repetition and was used to highlight the topic but not the key points.	Intonation was used effectively to keep audience interest and to emphasise the meaning of the topic, although the key points were not emphasised.	The only visual aids used were fingers to show the number of the several points. This was sufficient for the topic.	Top effectively maintained interest with the phrase: **"If you're ready, say I'm ready!"** Top stopped using any techniques to maintain audience interest during the content phase. This may have been due to time constraints.	Top kept checking audience agreement with the points he was making by asking **"Yes?"** but this did not check audience understanding. The opportunity to assess and treat this was missed due to running out of time.	Top's presentation would have benefited if he had related the topic to the audience in their working context.	This was not done although I am certain Top would have concluded if he hadn't run out of time. He is an experience d and trained presenter.	Top ran out of time which indicates that either the content was too great, or that the content should have been more effectively summarised throughout.	Top basically started off very well and got worse. This is an unfair reflection in some ways because I think this was mostly caused by the tight time limit. I am in no doubt that Top would have maintained the same professionalism throughout if he had not been caught out by the time limit.

Song	The introduction was basically just stating the topic. This could have been used to excite the audience about what is a relatively exciting topic.	No rhetorical questioning was used.	No other use of questioning was used.	This was not used but could have been used to highlight the key points.	Song's intonation was much like his usual spoken style. His presentation would have benefited from a 'performer' personality.	This was not used but could have been effective in getting the key points across.	Little interest shown. No observable techniques used.	This was not done.	This was not done. Song's presentation would have benefited if he had related the topic to the audience in their working context.	This was not done. A missed opportunity to repeat the key points, personalise, or check understanding	Song overran the time limit significantly. Content could have been better summarised.	Song's ability to present the SHEC Communication seems limited by his level of English fluency. Song could definitely benefit from utilising some of the techniques listed here.

Categor-ising	Understanding	Questioning	Emphasis	Audience Understanding	Time	
Comments and action by criterion	Introductions are very important and would be a great focus for a workshop. All presenters could benefit from some training on hooks and the need to plan these beforehand.	All presenters could benefit from some training on rhetorical questions and the need to plan these beforehand. I would also like to encourage the use of questioning throughout the presentations and at the end as a method of checking audience understanding.	All presenters need some work on this. This should be connected to the work I want to do on preparation of the key points – highlighting the key points on the SHEC Communication document. All participants could do with some focus on the identification of affordances for visual aids use, the variety of visual aids possible, common mistakes with visual aids, preparation of visual aids at the planning stage, and effective use of visual aids.	Training on this will require a change of mindset. A lot of students have a adopted the 'lecture' approach whereby the presenter only has to present the information and it is up to the audience to understand it or not. I would like to design a kind of workshop which incorporates skills practice but also encourages presenters to take on the responsibility of audience understanding.	This is not a skill which I think students require particular training on. It's a matter of practice – practice that they will receive while practising the other techniques listed here.	Overall I have identified some very useful areas of weakness which I can use to design future lesson content.

Conclusions, Reflections, and Future Directions

There are various aspects of my assessment instrument that I feel could still be improved. I approached this study with the ideal that my performance assessments would be criterion-referenced in order that I would be empowering my students to reach an actual, measurable standard of competence. In practice, I soon realised that due to the uniqueness of the task, my diagnosis of gaps in student competence would have to be based only on my own conception of a realistic target competence for my students because no external standard exists. This also meant that evaluations were norm-referenced in the sense that I was judging students' performances based on my notion of what they are capable of, which is "inappropriate for formative assessment because it legitimates the notion of a standards baseline which is subject to existential determination" (Sadler, 1989, p. 127). To counteract the norm-referenced orientation of assessing students against my own concept of a reasonable target standard of competence, I would have ideally preferred to include at least one more assessor to increase the objectivity of the generated data and achieve triangulation, as Allwright and Bailey (2004, p. 73) advised, "at least two perspectives are necessary if an accurate picture of a particular phenomenon is to be obtained." Unfortunately this was not possible in this instance. An additional weakness of the data generation was that starting the evaluations with a list of predetermined criteria meant that I was not receptive to aspects of students' strengths and weaknesses which were not included on the list. Ordinarily this would not be desirable for assessing student presentations but it may have been useful for diagnostic purposes.

Despite the criticisms mentioned above, there are aspects of the data generation that I am content with. I feel that the specificity of the criteria, basing the initial assessments on descriptors, and the repeated viewings of videoed presentations meant a thorough diagnosis of the gaps in each student's competence. I also feel that the thematic chart approach meant that more descriptive data was collected which led to more effective analysis and interpretation, and more specific diagnosis. Also, utilising the thematic chart during the ultimate stage of the data generation addressed concerns about norm-referencing to some extent, because the data became a great deal more descriptive and therefore more transparent. Comments, even if they are somewhat subjective, by nature provide the reader or analyst with more information than grades or band scores.

The most important conclusion I have drawn from this study is that teachers can work independently to diagnose their students' needs before tackling the task of addressing those needs. A thorough diagnosis increases the likelihood that the teacher can meet the students' specific requirements. I wanted to ensure that this study was informed by a basis of established research, and conducted in a manner which was as objective as possible. I conducted this research in a pragmatic manner, in essence just tackling each stage in order with very little ability to foresee the subsequent stage. Of utmost significance is the fact that I take data and conclusions away from this research that I will use to begin an action research project into using formative assessment to improve my students' proficiency on performance assessments. The areas of weakness identified here, will dictate the focus of future lessons and projects.

References

Allwright, D., & Bailey, K. M. (2004). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge: Cambridge University Press.

Biehler, R. F., & Snowman, J. (1997). *Psychology applied to teaching* (8th ed.). Boston: Houghton Mifflin Harcourt.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi: 10.1080/0969595980050102

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*, 393-415.

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 469-482). Malden: Blackwell.

McCullen, C. (1997). Presentation rubric Retrieved 13 January, 2012, from http://www.ncsu.edu/midlink/rub.pres.html

NCTE/IRA. (2004). Oral presentation rubric Retrieved 13 January, 2012, from http://www.readwritethink.org/files/resources/lesson_images/lesson416/OralRubric.pdf

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. doi: 10.1007/bf00117714

Simkins, M. (1999). Designing great rubrics Retrieved 13 January, 2012, from http://www.registereastconn.org/sblceastconn/greatrubrics.pdf

Spencer, L., Ritchie, J., & O'Connor, W. (2003). Analysis: Practices, principles and processes. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 199-218). London: Sage.

Swinton, L. (2012). Oral presentation rubric: How to get great presentation grades Retrieved 13 January, 2012, from http://www.mftrou.com/support-files/oral-presentation-rubric.pdf

Argument-Based Validity in Classroom and Program Contexts: Applications and Considerations

Justin Cubilo
cubiloju@hawaii.edu
University of Hawaii at Manoa

Central to determining the quality of any measure of learner ability is the determination of whether such measures provide a valid assessment of the abilities under question. The notion of what validity is and how to assess the validity of a given measure has undergone several changes over the past half century. Early conceptualizations of validity focused on the notions of criterion, content, and construct validity as more or less separate models. However, it has been recognized that criterion and content validity, while useful, are limited in what they can provide as supporting evidence for establishing validity since when they are used individually they only address a smaller portion of what needs to be considered for assessing the validity of a measure. This led some theorists such as Loevinger (1957) to suggest that criterion and content validities were simply parts of validation which fell under the umbrella of construct validation. Based on this view of validation, Messick (1989) proposed a unified model of validity, which included empirical methods for construct validation and consequences for test interpretation and use. At this time, Messick (p.13) defined validity as:

> An integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. [italics in original]

Thus, with his definition Messick removed the test itself from being the focus of validation and instead placed the focus on the score interpretation and use. This would ideally be accomplished through the construction of a logic-based validity argument by gathering the necessary evidence for and against the proposed interpretation or use of the test score and the inferences that are associated with these interpretations. Kane (2006) outlines such an argument-based approach, which is described below.

An Argument-Based Approach to Validity

According to Kane (2006), validation consists of two types of arguments, an interpretive argument and a validity argument. The interpretive argument is built upon a number of inferences and assumptions that are meant to justify score interpretation and use whereas the validity argument evaluates the interpretive argument in terms of how reasonable and coherent it is as well as how plausible the assumptions are (Cronbach, 1988). Development of such arguments requires the use of a clear structure on which the argument may be based. For this reason, those who work on developing interpretive and validity arguments (Kane, 2001; Mislevy, Steinberg, & Almond, 2003) base their arguments on Toulmin's (1958, 2003) framework for creating informal arguments, which essentially requires that a chain of reasoning be established that is able to build a case towards a final conclusion, which in this case would be to determine the plausibility and reasonableness of score interpretations and uses.

As is shown in Figure 1, Toulmin's (2003) argument structure is built on several components, which include the grounds, claim, warrant, backing, and rebuttal. As it relates to test score interpretation and use, the claim of an argument is the conclusion one draws about an individual based on test performance

whereas the grounds serve as the data or observations upon which the claim is based upon. For example, one may make the claim that an individual learning English has inadequate listening comprehension abilities for studying at an English medium university based on the grounds that they received a low score on a multiple-choice listening comprehension test consisting of a series of lectures utilizing academic vocabulary and structures. However, the inference linking the grounds to the claim is not given and therefore justification is needed in the form of a warrant (or assumption). The warrant in Toulmin's model is considered to be a rule, principle, or inference-license that is meant to provide justification for the inference connecting the grounds to the claim. Warrants in turn need backing (or evidence) which comes in the form of theories, research, data, and experience.

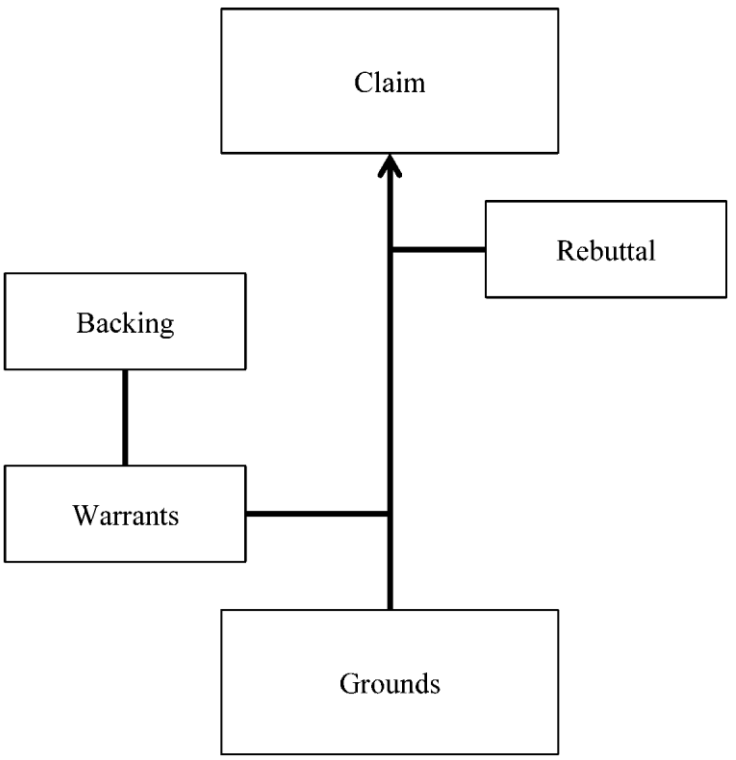

Figure 1. Model of Toulmin's (1958; 2003) argument structure.

In relation to the example provided above, the warrant justifying the inference between the grounds and the claim would be that performance on the listening comprehension tasks reflect relevant and necessary language abilities needed in an academic context. This warrant would then be supported by backing that might say that individuals with low-level listening ability generally have difficulty understanding academic words, making inferences or predictions from what a speaker has said, or poor knowledge of signal words and phrases meant to hint at main ideas or important points and that such deficiencies lead to poor performance in an academic English-speaking context. Finally, while warrants and backing justify the inferential link between the grounds and claim, rebuttal data can serve to weaken the initial argument by providing evidence or possible explanation which may call into question the warrant. Going back to the previous example, a possible rebuttal may be that several of the topics presented in the lectures may have been too technical or abstract, the vocabulary may have consisted primarily of less commonly or

frequently used academic vocabulary, or even that the audio quality was poor. Such data would serve to weaken the inference connecting the grounds and claim and would either have to be investigated further or accepted by the test developer with the knowledge that it places a limit on the argument. Thus, these components are all connected with each other and are essential for establishing an inferential connection between the claims and grounds.

In order to establish a connection between the claims and grounds, Kane (1992) stated that multiple inferences of different types must be used in a chain to connect observations and conclusions. Therefore, Kane, Crooks, and Cohen (1999) developed a three-bridge model for the three types of inferential bridges they thought were essential for linking arguments together in order to move from observation (i.e., the grounds) to score interpretation (i.e., the claim). Each inference is in turn based on a series of assumptions, each of which requires support. These three inferences were identified as evaluation, generalization, and extrapolation inferences. The evaluation inference refers to the score that is assigned to an individual's performance on a measure with the underlying assumption that appropriate criteria are used to score the performance, that they have been applied as planned, and that the conditions under which the performance took place match the intended score interpretation (Kane, 2002b; Kane, 2013; Kane et al., 1999). Following the evaluation inference, the generalization inference refers to the use of an observed score as a way of estimating future performance or scores of a test taker if given parallel tasks or test forms. Finally, following generalization is the extrapolation inference that refers to predictions of how the expected score is to be interpreted as an indication of performance and scores that the individual would receive in the target domain. An important assumption of extrapolation is that test tasks are authentic relative to tasks test takers would be expected to perform in the target domain.

In applying the bridge model to language testing, Chapelle, Enright, and Jamieson (2008) describe three further inferences in their validity argument for the TOEFL iBT that can be used to strengthen the connection between the grounds and claim and these are labeled as the explanation, domain description, and utilization inferences. The explanation inference describes the relationship between the observed test performance and a theoretical construct (e.g., a construct of second language listening). The domain description inference refers to a detailed description of the target domain and is meant to provide a link between performances in the target domain and observed performance on the test. Finally, the utilization inference provides the link between the target score that has been obtained for the test taker and the decisions that will be made about the test taker in relation to policy. Taken together, these six inferences along with their assumptions and support, which is obtained through a variety of methods, are able to provide a chain of arguments that can support the link between the grounds and claims of the overall validity argument. The types of evidence that can be collected as support for the assumptions in each of these inferences is manifold.

Applying the Argument-Based Validity Framework

Kane's framework (and its expansion by Chapelle) is a useful tool for considering the interpretations and uses of test scores. However, since it is slightly abstract in nature, it can be difficult for instructors and administrators to fully realize how to apply it to in their specific situations. In essence, to fully comprehend how this framework can be utilized within a particular situation, being able to see how evidence can be acquired to provide support for the different inferences within the model is necessary. This will ensure that teachers' and administrators' score interpretations and uses can be fully supported in their particular contexts. Below I discuss how teachers can gather evidence for some of the more relevant inferences for the classroom context.

One of the first things that instructors or administrators must do in creating a valid test for their classroom or for placement purposes is to adequately define the domain that they are attempting to assess. In order

to accomplish this, instructors and administrators can do several things. The first think that should be examined are the curricular and course objectives and student learning outcomes. These should be used to guide discussion regarding the content of the assessment and for determining the appropriate question formats for adequately assessing these outcomes and objectives. For instance, if students are expected to display appropriate pragmatic knowledge in making refusals at a certain level in a program or by the end of a course, a test should include a component that is meant to assess this ability and a role play or some other speaking activity may be more appropriate for assessing such knowledge rather than a multiple choice test. Additionally, it is important that the underlying trait, or construct, is appropriately defined so that educators can be absolutely sure that they are assessing what they wish to assess. Having a clearly defined construct will essentially provide stakeholders with a clear idea of what characteristics should be incorporated into an assessment meant to measure a given skill area.

Once constructs and outcomes and objectives have been clearly defined, teachers and administrators can proceed to develop appropriate tasks that are meant to target these aspects. This type of consideration is placed under the evaluation inference within an argument-based validity framework. This is essentially the time where a teacher can pilot the items of the test to gather evidence related to how they are working and to see how administrative conditions are affecting performance (Enright et al., 2008). In this way, teachers can revise their items or tasks by examining item facility and b-index or item discrimination values to see how items differentiate between learners on certain objectives. Furthermore, test adminstration characteristics can be investigated at this time to determine if such characteristics significantly affect performance in a positive or negative way. Examples of such characteristics would be to examine how the presence or absence of extra planning time for responses on speaking or writing tests affect overall performance or whether notetaking on a listening test significantly helps or hinders performance. This would also be an excellent time to get feedback from students who will be taking the test as they can tell you which of the administration conditions they prefer and how they relate to their affective state as this is something that could affect the overall relation between test performance and the construct that has been defined.

The generalization and explanation inferences come next in the argument-based framework, and it is here that some statistical evidence is required. Generalization in essence refers to the reliability of the assessment and whether the student would perform comparably well on future administrations of similar tests. This can be determined by calculating split-half reliability or the K-R20 or K-R21 values (where the test has been administered only once), by investigating test-retest reliability where a test is given twice and the results are correlated with each other, or by parallel forms reliability in which the test is correlated with another equivalent form targeting similar material (for more on the calculation of these coefficients, see Brown, 2005). Such information will provide the test designer with information on the amount of construct-irrelevant error that is present within the test so that they can determine how to proceed (often by either increasing the number of items found on a test or ensuring that test items are not ambiguous). Furthermore, teachers who are using tests as measures to determine who has mastered and who has not mastered content can evaluate the consistency of such determinations by using test dependability measures. One possibility for calculated this dependability index is to calculate the phi (lambda) coefficient which will provide the developer with information related to the dependability of a given cut score, taking into account the fact that some people pass a test to a greater extent than others. A discussion of how to calculate the actual coefficient is beyond the scope of this paper, but the reader is directed to Brown (2005) for his discussion of this topic.

Beyond the generalization inference, the explanation inference provides evidence to show that performance on a test is in line with the construct previously designed by the test developer. For instance, if the test that a teacher is developing is meant to assess achievement in meeting learning outcomes in an intermediate language skills classroom, the teacher can assess whether the test does indeed do this by

having beginning, intermediate, and advanced learners take the test in order to investigate differential item functioning. If the test and items function in accordance with what would be expected in relation to learning outcomes and objectives (i.e., intermediate students scoring significantly higher than beginning students and advanced students showing greater mastery of the outcomes and objectives than both intermediate and advanced groups) and the construct, then the teacher or administrator would have evidence to support the explanation inference. Furthermore, if the school has access to other measures of a similar skill (e.g., listening, speaking, writing, etc.) that they can have their students take, they can take results from these measures and correlate them with the measure they are developing in order to assess the test's convergent validity. This is effectively the correlation between two measures of the same or similar construct that use different methods (e.g., multiple choice and short answer questions or direct and semi-direct speaking assessments) (Crocker & Algina, 2008). Having a high correlation would show that a similar construct is being assessed and would lend credence to the support of the explanation inference. For a school or program environment, these two types of evidence would be good starting points for providing sufficient evidence for the explanation inference.

The final portion of constructing the validity argument requires providing support for inferences that focus on connecting test performance to performance and effects of score use outside of the actual test. The extrapolation inference is the first of these and can be supported through correlation studies (Kane, 2013). Whereas correlations in the explanation inference are used to provide evidence for the relationship between scores and the construct, correlations in the extrapolation inference are used to make connections to performance in the target domain and these correlations can be done with similar measures. They can also be correlated with course performance to ensure that there is a strong relationship between test performance and performance in the language or content classroom, which would indicate good fit for the test in relation to learning outcomes (which, by extension, would ideally be related to performance on real world tasks). For instance, if a teacher of English academic listening were seeking to obtain extrapolation for their test, they might correlate performance on their test with performance in lecture-style content courses and that performance in these content courses differs with each level of listening ability based on how many learning outcomes have been mastered by the student as displayed by the test score.

The utilization inference is the final inference in the argument-based validity framework and is often the inference that is addressed after a test has been developed and administered. This inference requires support for moving score interpretation to score use and requires examination of the consequences of the test and its effects on policy (Kane, 2002a; 2013). Bachman & Palmer (2010) outline a number of factors that are important in relation to score use in the decision-making process. Specifically, they mention that the consequences of the test should be beneficial for all stakeholders, reports should be clearly presented and easily interpretable, and the test has positive washback on instructional practice and learning. The utilization inference rests upon the assumptions that the consequences and decision-making process have been investigated in order to ensure that decisions and consequences equitable, scores are interpretable, and that instruction is positively affected by test use.

Teachers and administrators can do a number of things to gather evidence for this inference. First, washback studies can be conducted in which teaching is observed and learning is assessed. In this way it is possible to see whether course objectives are being targeted appropriately within the classroom and whether student learning as assessed by the new test is focusing on appropriate content and how this is related to topics covered in the classroom. Furthermore, stakeholder input from students and other teachers who may use the test would serve to be valuable in ensuring that the test is perceived as fitting with instruction and course objectives and that it is perceived as adequately assessing student mastery of specific learning outcomes. This type of feedback will serve to make for better score interpretations related to performance in the target domain. Finally, further investigations can be conducted in order to assess cutoff score determinations in order to make sure that such scores are appropriate for making decisions of

mastery versus non-mastery. This is especially important when achievement tests are used to determine if individuals have adequately learned the material to advance to a higher level in the language program. For a discussion of methods related to determining cut scores, the reader is directed to Brown (2005) and Fulcher (2010). All of this evidence will provide a clear and easy-to-follow blueprint for instructors to use so that they remember how to use their tests appropriately and how to interpret the scores.

Conclusion

Taken together, the evidence from each of the inferences mentioned above is put together into a single validity argument. The purpose of the validity argument is to determine whether the evidence that has been collected for each of the inferences is appropriate and actually supports the interpretations and uses of the tests either within a single classroom or program-wide. In order to condense the information, Table 1 summarizes the key points and sources of evidence for each inference.

Table 1
Summary of Inferences and Evidence Sources

Inference	Purpose	Evidence
Construct & Domain Definition	Describing and understanding the target domain (context) and skill to be measured to support intended interpretations	Literature Analysis Content Analysis (Examining program and class objectives and student learning outcomes)
Evaluation	Scores of observed performances are examined as measures of performance in the L2 ability. Highly relevant for determining score meaning	Item Analysis (Item Facility, B-Index, Item Discrimination) Stakeholder Opinions Examining effects of Administrative Conditions
Generalization	Ensuring that observed scores are consistent over future, parallel task versions so that expected scores can be estimated	Reliability Statistics (K-R20, K-R21, Split-Half Reliability, Phi (lambda))
Explanation	Determining that scores are related to the defined construct in a way that aligns with theory	Differential Item Functioning Studies Differential Group Studies Convergent validity
Extrapolation	Extension to performance outside of the test within the target domain	Correlation to performance in the target domain
Utilization	Moves from score interpretation to score use. Considers impact of test in relation to decision-making policies and curricular adaptation.	Stakeholder feedback related to perceptions of score interpretations Washback studies Cut Score Examination

It is recommended that inferences be addressed in the order that they are placed in the table as they will help to focus evidence for later inferences. While it is not always possible to gather evidence for all of these inferences within a given context, it is preferable to do so as this will only serve to provide stronger support for intended score interpretations and uses, which is what test developers, in both major testing companies and in classrooms, should be striving to do.

Bibliography

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). New York, NY: Routledge.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R. N., Mollaun, P., Nissan, S., . . . Schedl, M. (2008). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 97-143). New York, NY: Routledge.

Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. T. (2002a). Practice-based standard setting. *The Bar Examiner, 71*, 14-24.

Kane, M. T. (2002b). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 18*, 5-17.

Kane, M. T. (2006) *Validation*. In R. Brennan (Ed.), *Educational Measurement*, 4th ed. (pp. 17-64), Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3*, 635-694.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, 3rd ed. (pp. 13-103). New York: Macmillan.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge, UK: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument: Updated edition.* Cambridge, UK: Cambridge University Press.

Testing at CAL*: An interview with Dr Meg Malone

*The Center for Applied Linguistics, Washington D.C.

By Daniel Dunkley
Aichi Gakuin Unversity

Margaret E. Malone (Ph.D., Georgetown University) is Associate Vice President for World Languages and International Programs at the Center for Applied Linguistics. She has more than two decades of experience in language test development, materials development, delivery of professional development and teacher training through both online and face-to-face methods, data collection and survey research, and program evaluation.

DD: Dr. Malone, thank you for meeting me today. To begin, could you please introduce CAL?

MM: CAL is a small not-for-profit located in Washington D.C. We were established in 1959 by Charles Ferguson and the Ford Foundation, and our mission is to improve communication through better understanding of language and culture.

DD: What is your role at CAL?

MM: I work on a variety of testing projects. Right now I'm the associate vice-president for world languages and international programs.

DD: What are typical CAL activities, apart from testing?

MM: We conduct a lot of work with professional development for K-12 English Language teachers, and we also have a small division working with refugee and immigrant services. Back in the 80s we worked with refugees from Vietnam and we had an office in the Philippines at the time.

DD: Could you tell me about a recent project?

MM: One of my favorite recent projects resulted in a publication in *Language Testing*. It was a project to look at language assessment literacy among foreign (or world) language teachers in the United States. Conducting research with foreign language teachers in the US is more difficult than with language learners, simply because you can't get the number you need to have a publishable study.

At CAL we have a directory of foreign language tests, which we started updating again on line in 2005. We conducted focus groups with teachers and administrators to make the directory more useful. We put the search terms in so that individuals could actually find what they were looking for. We changed the name of it because we found that when we called it a *database* users thought they would go in and see a copy of the test. Of course test developers aren't going to leave copies of their test like that because it's a test security breach. The aim of the directory is to describe the test: what it does who it is for, how much it costs, where to get it and so on.

DD: How does the database relate to assessment literacy?

MM: In conducting the focus groups we found that there's a need to educate teachers and administrators about how to select tests. So we developed a tutorial to accompany it. We conducted research with teachers and administrators on what was needed for the tutorial. We also conducted research with the language testers and we found quite a bit of difference between what language testers thought was important and what teachers thought was important.

DD: Can you give a specific example?

MM: There was one question where we were asking about how a page looked. One language testing reviewer explained how this page was not relevant to the current view of a validity argument. So it seems that every question was interpreted as "What does this have to do with the technical aspects of testing?" On the other hand, teachers said "Keep it as short as possible"; they want explanations to be short and to the point.

DD: What happened after this survey?

MM: It was a US federally funded grant, like most of my work, and we made some recommendations to change the tutorial to make it teacher-friendly, but also to make it reflect current research in language testing. We aimed to give teachers accurate and current information, but not so much that you lose them.

DD: Do you know how many teachers are using this tutorial?

MM: There are several thousand "page-hits" a year. So there are several thousand who use the tutorial every year, and even more who use the directory of language tests. Part of the reasons behind the tutorial was to have information available so that teachers and administrators looking for a test could work through and decide what they needed on their own.

I've had calls from someone who says "Can I use this test?" and half way through the conversation I realize that the caller is talking about a test for high school, but they are teaching kindergarteners. So there's a real mismatch. We wanted to make something available all the time that could help people make good decisions about tests to use.

DD: How do people use the system?

MM: The directory and the tutorial go hand in hand. The idea behind the tutorial is that you conduct a needs assessment: What's your population? What language are you looking for? What are you trying to test? Then when you search for a test, you look at the test critically: Maybe this isn't the right one for me? In some languages there are so few tests available that users jump to conclusions. For example if you click on Arabic they may assume that every test is right for them. So, users may mistakenly choose a high school test when they need an Arabic for kindergarteners test.

DD: What's the relation of your tutorial to in-service teacher training?

MM: It's complementary. For example I've been teaching recently at the University of Maryland. I usually have my students conduct a search and write an essay on what they found, whether they think it's appropriate for the population, what's missing and what they think should be available.

DD: Does the tutorial work as a distance learning course in testing literacy?

MM: We have workshops with serving teachers at CAL. We actually tested our directory and tutorial with the course participants, to find out what teachers were looking for. So I think the tutorial is a nice part of this kind of event, but it wouldn't say it would be the whole thing.

DD: Will this site continue in the future?

MM: We hope so. We're applying for more funding to keep it going.

DD: Let's talk about your experience as a Peace Corps language training administrator. The need for quick language training must present special challenges.

MM: All Peace Corps volunteers, both currently and at the time I was there, from 1996 to 2000, have pre-service training: language training, technical training, health and safety training, and cross-cultural training.

DD: How about your role as a linguist?

MM: At the end of training, volunteers take an Oral Proficiency Interview to make sure that they have enough language to survive. For 30 years we've training locals to test the volunteers. My job was to train those testers, and to keep track of the scores that volunteers received. Those volunteers were tested at the end of pre-service training, sometimes after one year of service, and then at the end of service. We want to see if they are maintaining their scores, or going up or down. So I provided training courses for OP testers. I worked with about 60 countries over the four years I was there- about 150 languages.

DD: So, mostly languages you didn't speak yourself?

MM: It's actually very freeing to work with so many languages- you let go.

DD: What was your conclusion about immersion language training. Was it successful?

MM: The motivation for Peace Corps volunteers is very high. The classes are very small- three to five per class. There's a lot of differentiation of instruction, a lot of checking. The tests are aligned very closely with the curriculum. Many of the teachers were also testers, so they understood the goal that volunteers were working toward. They could use that to inform instructional decisions, and move volunteers around from group to group. I worked with one set of sites that achieved high proficiency after 12 weeks of training in a language they had never learned. So I worked with them for a couple of years, and developed a standard for what was a reasonable expectation after twelve weeks. This had an effect on the language training. That was very satisfying.

DD: So these were very special niche tests: ESP speaking only.

MM: We also used an Oral Proficiency Interviews, and the ACTFL guidelines that accompany them. There were small number of each tests, but many sessions. We conducted about 5,000 Language Proficiency Interviews per year.

DD: So you have a fascinating past; how about future projects?

MM: One project that's very important now is Language Resource Centers. There are 15 Centers that are designed to improve teaching and learning of foreign languages. Unfortunately in 2011 we were cut by 50 percent by the US Department of Education, so we're really working to try to maintain services in very tough fiscal times.

DD: What do the centers do specifically?

MM: The one I work on, called the National Capital Resource Center, is focused on language teacher education. For example we have an on-line course in the basics of language assessment literacy. It's not for credit. It's a five module course that teaches the basics of assessment. We also run an annual conference, the East Coast Association of Language Testers, which I and my college Paul Lowinky founded in 2002. In addition we continue to update the database of language tests, and continue to conduct research with teachers on what they need in a language assessment resource. We also work in teacher

training, consulting the professors who train the teachers, to find out what resources they need to help their students.

DD: Could you give me your thoughts on what the language testing community needs to work on in the next five years?

MM: I think we need to continue to encourage language assessment literacy. It's not just about understanding what assessment is, but about understanding what reasonable expectations are for students learning languages. Many administrators and parents have unrealistic expectations, either too high or too low, about what you can attain in a short period of time, and what's needed to get to that level.

One more thing that I'd like to see is a national study of foreign language outcomes. The last one was as long ago as 1965. So we don't know what our language majors have learned, or what K-12 students achieve in world languages. We need a study of what is going on nationally.

DD: What about types of test?

MM: There's definitely more computer-based testing with, for example, Parkin and Smarter Balanced; they are two organizations that are developing national tests in the core areas- reading and math. Then there's *Access,* which is a test used in 30 plus states to show that students are achieving the *no child left behind* goals. CAL is the test developer for that test, and we're offering a computer- based version of it. But it's really important that we also test speaking. We've been using the computer to elicit the language, but it has to be rated by humans.

DD: That must make it an expensive project.

MM: True but it's more economical than sending out examiners to do oral interview tests. Also, it's more reliable, because you're getting responses to the same tasks.

DD: Well, I hope that as a result of your and CAL's efforts in assessment literacy, Language Resource Centers, outcome tests and computer based tests, foreign language teaching improves across the US. Thank you Dr Malone.

Bibliography

Center for Applied Linguistics. (2007). *Understanding assessment: A guide for foreign language educators*. Retrieved 1 June, 2014, from www.cal.org/flad/tutorial.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329-344. doi: 10.1177/0265532213480129

Editors note: At time of publication we are still waiting to hear final confirmation from Dr Malone that this version of the interview contains no factual errors, so the online version may be updated to correct any errors. Any corrections will be footnoted to avoid confustion.

Questions and answers about language testing statistics:

Differences in how norm-referenced and criterion-referenced tests are developed and validated?

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

What are the major differences between norm-referenced and criterion-referenced tests? How can these two tests be best developed and validated? [Submitted by a participant in the Kuroshio (Aloha Friday) Seminar that Kimi Kondo-Brown and I conducted on May 23, 2014 at the Bunkyo Civic Center in Tokyo]

Answer:

I have discussed the major differences between norm-referenced and criterion-referenced tests in a number of places (most recently in Brown, 2012a). So I will only touch on those differences briefly here. I have also explained at length the different strategies that should be applied in developing and validating the two families of tests in a number of places. However, I have never summarized those different strategies side-by-side in one short and straightforward article. I will attempt to do just that here by addressing the following sub-questions: What are the differences between the norm-referenced and criterion-referenced families of tests? What strategies are used to develop and validate NRTs and CRTs? What are the differences in NRT and CRT development and validation strategies?

What are the Differences Between the Norm-Referenced and Criterion-Referenced Families of Tests?

Norm-referenced tests (NRTs, sometimes referred to as *standardized tests*) and criterion-referenced tests[1] (CRTs, also known as *classroom tests*) are two families of tests that are distinguished most clearly in terms of the ways scores are interpreted, the purposes of the tests, levels of specificity, the distributions of scores, the structures of the tests, and what we want the students to know in advance. In more detail, the two types of tests differ in:

- *The ways scores are interpreted* differ is that NRTs are designed to compare the performances of students to one another in relative terms, while CRTs are built to identify the amount or percent of the material each examinee knows or can do in absolute terms.
- *The purposes of the tests* also differ with NRTs primarily designed to spread examinees out on a continuum of general abilities so examinees' performances can be compared to each other

[1] Note that, since the question addressed to this column was clearly written by a person interested in testing, but primarily a teacher, the types of CRTs I am referring to here are not the formal subcategory of CRTs known as domain-referenced tests (which tend to be large scale), but rather those CRTs used by teachers on a more focused classroom level.

(usually with standardized scores), while CRTs are designed to assess the amount of material that the examinees know or can do (usually expressed in percentages).

- *Levels of specificity* are necessarily different with NRTs tending to measure very general language abilities (for proficiency or placement purposes), while CRTs usually focus on specific, well-defined (and usually objectives-based) language knowledges or skills (for diagnostic or achievement purposes).
- *The distributions of scores* also differ in that, ideally, NRT scores are normally distributed (indeed items are selected to ensure this is the case), while CRT scores ideally would produce quite different distributions at different times in the learning process: with students scoring very low in a positively skewed distribution at the beginning of a course on a diagnostic CRT (indicating that they needed to learn the material) and students scoring generally high in a negatively skewed distribution at the end of the course on an achievement CRT (indicating that most of them mastered the material; indeed, in the unlikely event that all students master all the material, they should all score 100%).
- *The structures of the tests* also differ with NRTs tending to have many items with a few long subtests (e.g., listening, grammar, reading, etc.) each of which has diverse item content, while CRTs are typically built around numerous, short subtests that contain well-defined and similar items in each.
- *What we want the students to know in advance* of the test differs in that, for NRTs, security is usually an important issue because we do *not* want examinees to know the content of the test items, while for CRTs, we teach the content of the course and want the students to study that content, so we tell them what to study, and we test that content. If they know the content, they should succeed.

What Strategies Are Used to Develop and Validate NRTs and CRTs?

Table 1 summarizes the strategies used to develop NRTs and CRTs in two separate columns. I hope that this table is clear without any direct explanation. Nonetheless, some discussion of the differences between NRT and CRT development strategies will be provided below.

Table 1
Strategies Used to Develop NRTs and CRTs

Steps	***NRT (Standardized)***	***CRT (Classroom)***
1. Plan test	Plan based on test specification/blueprint and general item specifications.	Plan with course objectives developed and in hand; when possible, using item specifications will help.
2. Create items	Create a large pool of items at about the right level of difficulty in the general area being tested (e.g., reading comprehension).	Create about 10 items that measure what the students should be able to do on each of the course objectives (say objectives 1-9) at the end of the course; divide the items into two forms of the test, say forms A and B such that there are about 5 items on each test for each of the 9 objectives/subtests.
3. Edit items	Use item writing guidelines like those found in Brown (2005, Chapter 3) to carefully proofread and improve all items.	Use item writing guidelines like those found in Brown & Hudson (2002, Chapter 3) to proofread and improve all items. Perform item congruence and applicability analysis (as described in Brown & Hudson, 2002, pp. 98-100) to make sure items match objectives.

4. Pilot items	Pilot the items *with a single large group of examinees* that has the same characteristics and range of abilities as the examinees in the ultimate test group (e.g., if the test is being developed for proficiency purposes, pilot it with a large group of students ranging from near-zero English to near-native; if the test is for placement purposes at a specific institution, the test should be piloted with examinees in the narrower range of abilities found there).	Ideally, pilot the two forms *at the beginning of the course* as diagnostic tests (with half of the students randomly selected to take each form); score and give the students diagnostic feedback objective-by-objective based on the subtests. Then, administer the same tests *at the end of the course* as achievement tests such that students who took Form A at the beginning take Form B at the end, and vice versa; include the scores in the students' grades, but keep the tests for further analysis.
5. Analyze items	Calculate *item facility* (*IF* = the proportion of examinees who answered each item correctly) and *item discrimination* indexes (*ID* = proportion of examinees in the upper third on the whole test who answered each item correctly minus the proportion in the lower third) (see Brown, 2005, pp. 66-76).	Calculate *difference indexes* (*DI* = proportion of students who answered each item correctly at the end of the course minus the proportion at the beginning) and *B indexes* (*BI* = proportion of those examinees who passed the whole test that answered each item correctly minus the proportion of correct answers for those students who failed) (see Brown, 2005, pp. 76-84, or Brown & Hudson, 2002, pp. 118-148).
6. Select items	Revise the test by selecting those items with the highest *ID* values while keeping an eye on the *IF* values to adjust the difficulty of the test up or down as necessary.	Revise the test by selecting those items with the highest *DI* values within each objective/subtest (perhaps the best 3 out of 5). If *DI* values are not available, select the highest *BI* values in each objective/subtest (again, perhaps the best 3 out of 5).
7. Revise test	Create a new, shorter, more efficient revised test based on the item analyses and selection in Steps 5 and 6 for future proficiency or placement purposes.	Create new, shorter, more efficient, revised Forms A and B based on the item analyses and selection in Steps 5 and 6 for future use as diagnostic and achievement tests.

Table 2 summarizes the strategies used to validate NRTs and CRTs in two separate columns. Again, this table should stand alone as a summary, but further discussion will be provided in the next section.

Table 2
Strategies Used to Validate NRTs and CRTs

Steps	***NRT (Standardized)***	***CRT (Classroom)***
8. Examine consistency	Study the *reliability* of scores by using test-retest, parallel forms, or internal consistency strategies–the most commonly applied internal consistency estimates are Cronbach alpha, K-R20 or K-R21 (for full explanations of all these reliability strategies, see Bachman, 2004, pp. 153-191; Brown, 2005, pp. 169-198; Brown, 2013a).	Study the *dependability* of scores by using threshold loss agreement (agreement or kappa), squared error loss (Φ_λ), or domain score dependability (Φ) strategies. If resources are limited as in most classroom settings, teachers can use the K-R21 reliability statistic as a conservative estimate of Φ mentioned above (for full explanations of these dependability strategies, see Bachman, 2004; pp. 192-205; Brown, 2005, pp. 199-219; Brown, 2013b).
9. Examine validity	Use evidential strategies, which include the traditional content, construct, and criterion-related validity strategies. Also use the more recently developed consequential strategies including examination of the values implications and social consequences of score interpretations and uses (see Bachman, 2004, pp. 257-293; Brown, 2005, pp. 220-248).	Use the only evidential strategy that typically makes sense for CRTs, which is the traditional content validity approach. Teachers may also want to use the more recently developed consequential strategies that take into the account values implications that they are expressing by the choices they make in test design as well as the social consequences of their score interpretations and uses (see Brown, 2012b; Brown & Hudson, 2002, pp. 212-268).

What are the Differences in NRT and CRT Development and Validation Strategies?

Careful examination of Table 1 will reveal key differences between NRT and CRT development strategies. In Step 1, the primary difference in *test planning* is that CRTs are more specific and objectives-based,

while NRTs are more general. In Step 2, the difference in *creating items* is that a more general pool of items is developed for NRTs, but smaller, more specific item pools are created for each objective/subtest in CRTs. In Step 3, *editing items* includes using item guidelines for both types of tests, but item congruence and applicability analyses are key to CRT development. In Step 4, the key difference in *piloting items* is that NRTs can be piloted in one shot and must include the whole range of abilities being tested, while CRTs are best piloted at the beginning and end of appropriate instruction and should focus only on what is being taught. In Step 5, the key difference is that *analyzing items* for NRTs is based on *ID*, and *IF* (in that order), while ideally, CRT item analysis is based on *DI*, but in a pinch can be based on *BI*. In Step 6, the key difference in *selecting items* is that, for NRTs, it is based on the highest *ID*s, and then on *IF* (to adjust test difficulty), while ideally CRT item selection is based on the highest *DI*s, but in a pinch on the highest *BI* values. In Step 7, the prime difference in *test revising* is that the ultimate product for NRTs is typically one large general test (or sometimes large subtests like grammar, listening, reading, etc.), but for CRTs, the resulting product is usually a collection of small, focused, objectives-based subtests, ideally in two forms

Table 2 reveals key differences between NRT and CRT validation strategies. In Step 8, the NRT *reliability* practices listed in the table are those laid out and explained for NRTs in most language testing (or more general testing) books. For CRTs, the *dependability* procedures shown in the table can clearly become quite elaborate. However, teachers need only address the common sense questions of whether the scores on their tests are consistent, fair, and consistently represent the knowledge and abilities of all students. If resources are limited as is the case in most classroom settings, teachers can use the K-R21 reliability estimate as a conservative estimate of domain-score dependability (Φ) referred to in the table (see argument for this strategy in Brown, 2005, p. 209).

In Step 9, the *validity practices* for NRTs listed here are also those laid out and explained for NRTs in most language testing (or more general testing) books including evidential strategies like content, construct, and criterion-related validity strategies and consequential strategies examining values implications and social consequences. For CRTs, the content validity approach listed in the table is the only one that always makes sense; it involves systematically analyzing and assessing the degree to which test items are measuring what the teacher is claiming to test, often by laying out the test items side-by-side with the course objectives (and with the teaching materials nearby for reference) and systematically comparing items to objectives. There are three key questions that teachers may want to consider in this regard (note that these questions and those in the next paragraph are adapted from and explained more fully in Brown, 2012b, 2013c):

1. Does the content of my test match the objectives of the class and the material covered?
2. Do my course objectives meet the needs of the students?
3. Do my tests show that my students are learning something in my course?

Teachers may also want to consider the *values implications* of their testing, scoring, and decision making by addressing some or all of the following questions: How do the learning/teaching values that underlie my test design, the resulting scores, and the decisions based on them match my beliefs and values? The beliefs and values of my students? Their parents? My colleagues? My boss? Etc.? Teachers may also want to think about the *social consequences* of their scores and decisions by addressing the following questions: What will happen to my students as a consequence of the decisions I make based on these test scores? Is this a small-stakes decision that is only a small part of a course grade, or will this test have larger consequences for students (e.g., determine whether or not the student passes the course, graduates with a diploma, etc.)?

Conclusion

In answering the question posed at the top of this column, length restrictions limited me to summarizing the differences in characteristics, development steps, and validation strategies used for NRTs and CRTs. I hope that this overview will nonetheless prove useful to readers and that anyone who wants more in-depth coverage of any aspect of these differences will be able to use the citations and references provided here to continue exploring these and related topics. I especially hope that this explanation will help practicing language teachers realize that most of the testing they do in the classroom ought to be CRT and that this column along with Brown, 2013c (which discusses solutions to problems that teachers often have with their classroom testing) will help them do a better job of assessing their students.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown J. D. (2012a). Choosing the right type of assessment. In C. Coombe, S. J. Stoynoff, P. Davidson, & B. O'Sullivan (Eds.). *The Cambridge guide to second language assessment* (pp. 133-139). Cambridge: Cambridge University.

Brown, J. D. (2012b). What teachers need to know about test analysis. In C. Coombe, S. J. Stoynoff, P. Davidson, & B. O'Sullivan (Eds.), *The Cambridge guide to language assessment* (pp. 105-112). Cambridge, Cambridge University.

Brown, J. D. (2013a). Classical theory reliability. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1165-1181). Oxford, UK: Wiley-Blackwell.

Brown, J. D. (2013b). Score dependability and decision consistency. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1182-1206). Oxford, UK: Wiley-Blackwell.

Brown, J. D. (2013c). Statistics Corner. Questions and answers about language testing statistics: Solutions to problems teachers have with classroom testing. *Shiken Research Bulletin*, *17*(2), 27-33. Also retrieved from the World Wide Web at http://teval.jalt.org/sites/teval.jalt.org/files/SRB-17-2-Brown-StatCorner.pdf

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Where to Submit Questions:

Please submit questions for this column to the following e-mail or snail-mail addresses: *brownj@hawaii.edu*. Your question can remain anonymous if you so desire.

JD Brown
Department of Second Language Studies
University of Hawai'i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA

Members' experiences and questions about testing and assessment

How to make a judging plan for rated tests?

Jeffrey Durand

Testing situation

A few years ago, I had to put together a speaking test for all the students (about 2,000) at my university. About 60 teachers were available to rate students, who were tested in groups of four. Two teachers worked together to rate all the students in each group. In speaking tests, the raters are often not equally strict (some tend to give slightly higher scores than others), and on occasion may give an unusually high or low score. These problems can be discovered by using software like *Facets* (Linacre, 2012), and scores can be adjusted or students can be retested. To do this, however, there needs to be a way to know how strict each rater is in comparison to others. This can only be done if all the raters (and tasks and prompts) are connected together in what is called a judging plan (Linacre, 1997; Sick, 2013).

I found a pretty good judging plan while observing a colleague's speaking class. The instructor put students into two concentric circles, with equal numbers of students in each circle. A student in the outer circle worked with a partner from the inner circle. After a period of time, the students in the outer circle all rotated one place around the circle to talk with the next student in the inner circle. This created a regular ring lattice in which each student could be connected to all the others. Figure 1 shows a regular ring lattice with 16 raters (the blue diamonds), each with three partners (connected by straight lines). A slightly larger version of this method seemed to provide exactly what I needed for the raters. It also fit the testing location, which took place on two floors of a building that has stairwells at each end. The raters could quickly and easily move between rooms. After the judging plan was set, it was easy to randomly assign students to each room at a certain time.

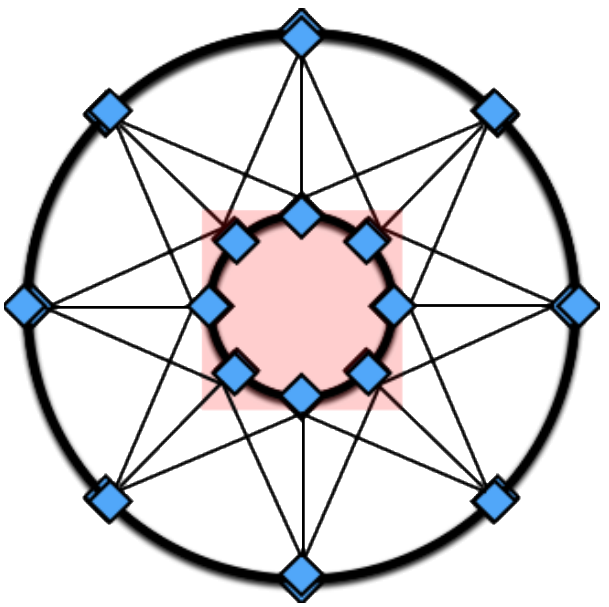

Figure 1. Judging plan

Questions

I have three questions about this judging plan.

1. How often should raters rotate, or for how many sessions should raters work together? Is it better to have raters work together for many sessions so that we are more confident about how strict they are in comparison to each other? Or should raters be rotated more often so that there are direct comparisons of strictness with more raters? Given that there is (thankfully) a limit to how many students an instructor is asked to rate, is there an optimal balance between rotating frequently and working with the same partner for a number of sessions?
2. Are there any other (better) ways of making a judging plan? For example, are there advantages of using three raters for each session or having an independent, trusted rater join random sessions? In your experience, what have been good (or not so good) ways of making judging plans?
3. Are there any questions that I have not considered that might be equally or even more important?

Do you have any real-life experience with judging plans or tests in which students are rated? Please share what you can!

References

Linacre, J. M. (1997). Judging plans and Facets. *MESA Research Note #3*, retrieved May 30, 2014, from http://www.rasch.org/rn3.htm.

Linacre, J. M. (2012). Facets (Version 3.70) [Computer Software]. Chicago: Winsteps.com.

Sick, J. (2013). Rasch measurement in language education part 7: Judging plans and disjoint subsets. *Shiken Research Bulletin, 17,* 27-31.

Where to Submit Questions:

Please send your responses to this question, as well as details about your own tests, to: tevalpublications@gmail.com

This section is a place for you, our readers, to share your experience with tests and to ask each other for advice. What you have learned can be a great help to others, both in the answers that you share and in the questions that you ask. When you submit your own questions about a test, remember to include a little background about it.