Meaningful Measurement: Teachers and Test Making

Jay Wong
Temple University
Takayuki Okazaki
Kyoto Sangyo University
Andrew Atkins
Shiga University

Reference Data:

Wong, J., Okazaki, T., & Atkins, A. (2012). Meaningful measurement: Teachers and test making. In A. Stewart & N. Sonda (Eds.), *JALT2011 Conference Proceedings*. Tokyo: JALT.

In this paper we describe teacher-led efforts to develop and revise a set of large-scale, university-wide tests for an English language program at a university in Western Japan. We hoped through this work to create a common measurement scale so that student performance on different versions of the tests could reliably be compared. We discuss the challenges of implementing teacher-managed test development in our work context and raise concerns about the lack of attention paid to construct mapping in the initial stages of the test development cycle. Using data from the reading section of our test, we show how we wrote and piloted test passages and items which gave us better information for student placement and end-of-term assessment. Also, we explain how we used Rasch modeling to determine how test items performed. In addition to statistical analysis of test items, we also suggest that qualitative student feedback be used when considering items for test inclusion or exclusion. Finally, we conclude that ideal test construction is an iterative and time-consuming process that offers the most promise to institutions willing to take a long-term perspective. With proper institutional support teachers can play a leading role in test development.

本稿は、西日本のある大学英語プログラムにおいて、教員が担当した大学全体で実施されるテストの開発と改善を報告するものである。まずある教育環境において教員がテスト開発を行う際に、影響を与える要因を検討し、テスト開発サイクルの第一段階で構成概念マッピングへの意識が希薄であったことを問題視した。次にリーディングセクションのテストデータを用いて、学生のレベル分けや学期末評価により正確な情報を与えるパッセージや問題の作成方法やパイロット法を示した。さらにテスト項目がどう機能したかを見出すため、どのようにラッシュモデルを用いたか解説した。統計的分析に加え、質的データとして学生からのフィードバックが項目決定に有効であることを提案した。最後に理想的なテスト構築は、時間と反復を要するプロセスであり、長期的視点を持つ教育機関が有望であると結論づけた。教育機関からの適切な支援のもと、教員はテスト開発にあたり先導する役割を担うことができる。

University language program administrators and teachers in Japan who aim to separate students into English ability groupings via placement testing face a number of significant challenges. For institutions with large numbers of students, test logistics are often difficult to manage as competition for lecture halls and classrooms at the beginning of the school term can be fierce. In addition, tests must be administered, scored, analyzed and interpreted in a very short time window before instruction begins so that placement into classes can be accomplished before instruction starts.

Even when testing is optimally organized, larger concerns with test validity and reliability remain. Commercially produced and validated tests such as the Secondary Level English

Proficiency test (SLEP), Test of English for International Communication (TOEIC), TOEIC Bridge, and Michigan Test of English Language Proficiency (MTELP) do exist, but for programs with limited resources costs can be prohibitive. Moreover, the sensitivity of such tests to place Japanese university students of English into multiple ability levels within a single institution may be somewhat problematic due to the normative nature of university admissions in Japan. Students in Japan tend to seek admission to universities based on their personal *hensachi*, a ubiquitous, but somewhat dubious standardized academic ranking score which classifies candidates on a 20-80 scale. At the same time, universities target students for admission using the same ranking system. A discussion of how the *hensachi* scores are produced and distributed is beyond the scope of this paper, but the relevant point here is that, practically speaking, both universities and candidates essentially target each other. One outcome of this process is that incoming cohorts of students tend to be drawn from generally more homogeneous academic ability groups, even if the degree of statistical validity of the ranking criteria remains questionable. In such a context commercially produced instruments may fail to distinguish student ability well enough for placement decisions to be trustworthy.

The alternative to off-the-shelf tests is logically to develop tests locally, specifically for the target population in question. Some obvious advantages to this solution include greater control of test content, cost control, and potentially improved test validity. In addition, such tests, properly validated, are more likely to produce better placement which can benefit the curriculum and instructional components of a language program. Understandably, many program administrators may hesitate to move in this direction leery of the investment in time, technical expertise and human resources required to oversee such a project and bring it to fruition. In this paper, we would like to show that teachers can take a leadership role in test development by sharing our own experience working with a university-wide test development project. We would like to suggest that teacher participation in such a development process can lead to better tests and better teachers.

Getting Started: Our Working Context

The setting for the test-development work described in this paper was a private university in Western Japan. The authors were hired as fixed-term contract teachers in the General Education division of the school and were asked to take on a testing project aimed at developing a set of three English proficiency tests. The first test form served as a placement test for all first year students and was administered at the start of the school year prior to instruction. The second test form was given at the end of the first semester and the third test was administered at the end of the school year. All three test forms were identical in format and had the same number and type of items. Test outcomes from each test had been analyzed by our senior colleagues using WINSTEPS Rasch measurement software (Linacre, 2011) and some initial attempts to weight test items based on their item difficulty had been made by these senior colleagues, but no formal attempt at scaling the three test forms had been attempted prior to our assignment to this testing project. In addition, we were told that the item weighting process used to calculate student scores on each test form had been somewhat ad hoc in nature, influenced by both the item difficulty statistics and the test makers' general impressions of the ability levels of the target test population. We understood that there was a desire to make the scoring process more transparent and trustworthy and given this general orientation we also understood that our primary task should be to analyze how well the tests were functioning statistically and to work on creating a common rating scale for all three instruments so that test-to-test results could be compared.

The first challenge we had to face was figuring out a way to proceed. Although we and our colleagues in the General English

program were well versed in creating classroom-based tests, few of us had any explicit training or experience in developing large-scale instruments. Understandably, without formal training in test development it was difficult to establish a clear road map to work from and it took us some time before we were able to create a common framework and vocabulary to discuss our testing work with each other and our colleagues.

In the following section we describe the context where we work, who our students are, who we are, what we were asked to do and why we were asked to do it. We also describe the historical background to the development of these tests.

Next, we discuss how we approached the process of reworking the test materials in a systematic way so that the equating of the three parallel test forms we inherited could be started and a common measurement scale for all three tests could be established. And finally, we raise some conceptual and procedural concerns about the lack of attention paid to construct mapping in our test development process.

Participants

The participants in our study include all first-year undergraduate students at our university not majoring in English or Cultural Studies. Their majors are Economics, Law, Business Administration, Science, Engineering, Computer Science, Life Science and Foreign Languages. In the school year of 2009, roughly 2,500 students were enrolled in Oral Communication (OC) or Reading Skills (RS) classes (See Table 1). There were also quite a few students taking both OC and RS classes, though due to student attrition and privacy concerns we, unfortunately, cannot report the exact number of students who completed both courses. As with most undergraduate students in Japan most of our students have had at least six years of compulsory English classes at the secondary level prior to entering university. Here, it is important to note that for most of our students, university English classes are *general education* courses required for graduation, but not highly valued as a top priority in their undergraduate careers.

Table 1. 2009 School Year Students and Classes

	Number of Students	Number of Classes
Oral Communication class	1645	75
Reading Skills class	1233	62

Testing Committee

As members of the Testing Committee, the authors worked at the aforementioned university as limited-term contract teachers who can only work at the university for a maximum of four years. We were each assigned to teach ten classes and be part of several committees, one of which was the Testing Committee. As noted above, in the Committee we inherited three sets of tests that were parallel in form and designed to a large extent to conform to old entrance exam item types. In other words, each test had the same number of items and the same types of items, but no statistical analysis had been done to examine the relative difficulty of each test and, consequently, there was no real basis to compare how well the tests had performed. So, we were asked to improve the quality of measurement for placement testing so that we could place our students more accurately into the five levels dictated by the university. We were also told that the three tests needed to be equated so that we could use them to measure students' improvement/achievement over the year they were studying English. We were asked to use the Rasch measurement model to help us evaluate the performance of these test forms. Briefly put, the Rasch model is a probabilistic

psychometric model that uses one test parameter, the difficulty of the items on a test, to estimate the relative ability of the people who take that test. Its principle advantage in our case was that it provided output in the form of ordinal data which is obviously quite helpful for scaling. Additionally, unlike more conventional statistical approaches, such a measurement model allowed us to analyze our test item difficulty data independently of the sample populations who were taking the tests.

The three test forms were administered at regular intervals in the first academic year. The first was used as a placement test in the week prior to instruction. The second test was treated as an end-of-first semester exam, while the third test served as a final exam at the end of one full academic year. Results from each of the latter two test forms were then used to determine 30% of students' grades. Despite the fundamental difference between placement and achievement objectives we were nonetheless asked to improve each test in the hope that the instruments could, with significant revision, evolve into viable achievement tests in the future. Still, the textbooks and curricula in each of our General English program classes at the time of writing, continues to vary significantly from teacher to teacher and level to level. Consequently, our second and third tests do not really function as achievement tests, but rather rough proficiency tests. We do, of course, share the hope that these tests reflect our students' proficiency gains over time.

History of the Tests

In the past, the English program at our university had no centralized test, no level-based classes, or required texts and a great deal of freedom was given to teachers. However, without having program-wide commonalities, the teaching content and learning outcome in General English courses varied tremendously. For instance, while some literature experts would teach interpretation of Shakespeare, others would teach how to improve their TOEIC score and others would teach how to make the best presentation. To centralize the content, people in charge of the program decided that they needed to control the curriculum and introduce the textbook list, the end-of-term test and placement test. It was hoped that these tests would have some "wash-back" effect, a beneficial influence on what and how teachers teach their classes.

A Model for Test Construction

After some early haphazard efforts to address wording, format, and clarity issues in our test forms, it quickly became apparent that as testing committee members, we needed to know more about how the items on all of the test forms were functioning. We realized that although we had a general picture of relative item difficulty on each of our tests, we lacked a clear understanding of what made the items difficult or easy and no data about the relative difficulty of the items between the three test forms. We also learned that a detailed analysis of test item performance still needed to be done. So we looked to the reference literature for some guidance on how to proceed.

After reading several introductory, but very helpful texts on testing and test development such as McNamara (1996), Fulcher and Davidson (2007), and Wilson (2005), we realized that a common and consistent theme amongst them all was a strong and consistent emphasis on the need to carefully consider the dimensions of the underlying construct to be measured as well as the social implications of how a test will be used. We use the term "construct" here to refer to the underlying characteristic being measured by our tests, which we may assume to be some aspect or combination of elements that contribute to our general perception of language proficiency.

We settled on Wilson's (2005) "Four building blocks" model of test development (see Figure 1) for its simplicity and con-

ceptual clarity. This model suggests an ideal work flow from construct mapping, to item generation, through to item scoring, and finally to the selection of a measurement model to analyze and interpret the test results. Wilson emphasizes the directional workflow of the model, but points out that at each stage the process is recursive with the direction of test development cycling back to earlier stages at any point in the process.

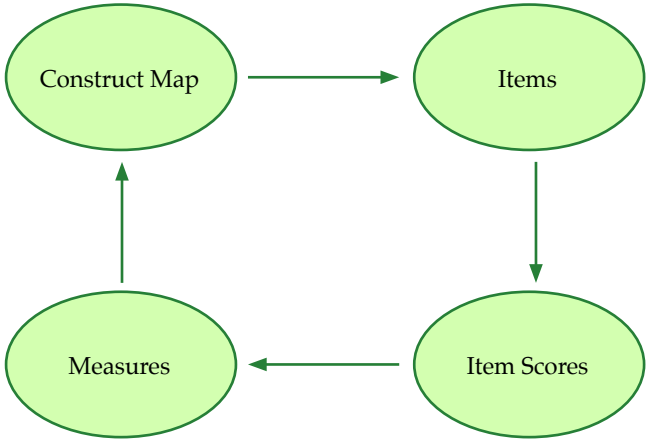

Figure 1. Wilson's (2005) "Four Building Blocks" model for test development.

Conceptual and Practical Problems

With a theoretical framework to orient us, we tried to make sense of the data we had collected. As we struggled to improve the clarity and performance of our test items we noticed how difficult it was for us as teachers to reconcile the statistical item difficulty of our test items with our intuitive sense of qualitative differences in our students' language proficiency. We noticed, for example, that sometimes items that discriminated well on our tests could be questions that appeared to assess eclectic, seemingly arbitrary grammatical knowledge while items that made intuitive sense to us as teachers were clustered around the person ability mean and thus contributed little to separate our students into ability groups. Rasch analysis software maps the relative ability of test takers onto a unidimensional scale and produces a *Wright map*, a visual distribution of their ability measures placed on an ordinal scale (see Figure 2).

Simply put, these person ability measures tell us how good students are at something, while controlling for the varying degrees of difficulty in the test items. So the dispersion and mean of this distribution tells us something about both student ability and test item performance. One major cause for concern in our case was the tendency for too many of our test items to cluster around similar points in the person distribution such as the person mean. If we examine Figure 2, we see that items16, 20, 3 and 8 all appear on the same line. Reading to the left of the vertical dotted line we can see that these items fall approximately in the center of the distribution of people. We can interpret this data as indicating that persons of approximately average ability are equally likely to get these items correct as to miss them. The meeting points of item difficulty and personal ability parameters on this kind of map are significant because they tell us both about the relative difficulty of the test items and the relative ability of the test takers. In Figure 2, we can see that most of our population remains clustered in one big bunch around the person mean and that our test items do not break up this cluster very well. Item redundancy is evident when multiple items appear on the same line. Finally, a large chunk of our test items fall far below the person ability mean indicating that they were much too easy for our population. In fact 14 out of 32 test items in this version of the test are not really contributing useful information about the ability levels of our students!

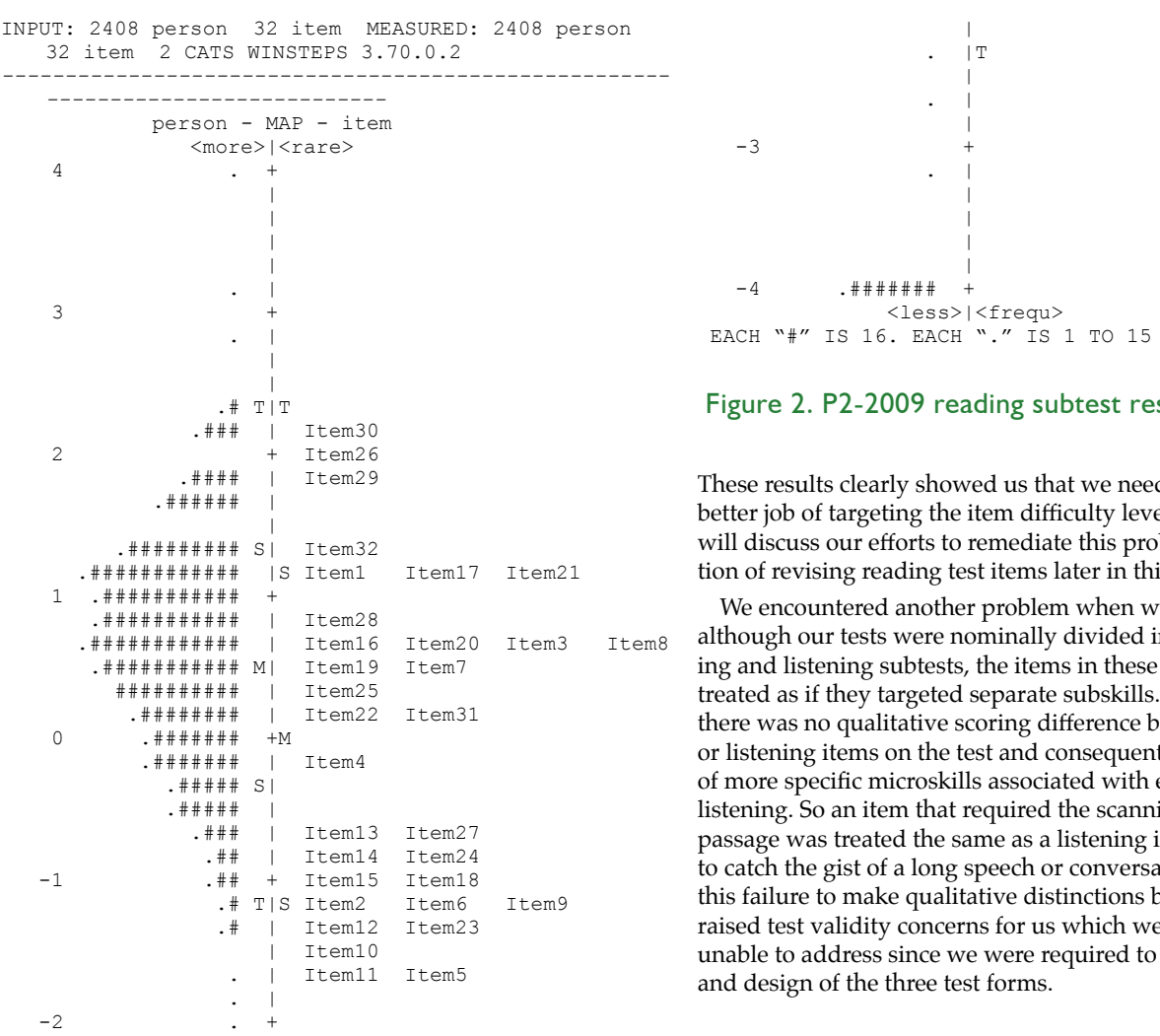

Figure 2. P2-2009 reading subtest results Wright map

These results clearly showed us that we needed to do a much better job of targeting the item difficulty levels on the tests. We will discuss our efforts to remediate this problem in our description of revising reading test items later in this paper.

We encountered another problem when we discovered that, although our tests were nominally divided into distinct reading and listening subtests, the items in these sections were not treated as if they targeted separate subskills. In other words, there was no qualitative scoring difference between reading or listening items on the test and consequently no mapping of more specific microskills associated with either reading or listening. So an item that required the scanning of a reading passage was treated the same as a listening item asking students to catch the gist of a long speech or conversation. Predictably, this failure to make qualitative distinctions between item types raised test validity concerns for us which we were ultimately unable to address since we were required to maintain the format and design of the three test forms.

We recognized that the lack of discrete test specification was conceptually problematic, but still faced the task of improving the measurement functioning of the tests. We felt that significant progress could still be made on this front by revising or replacing poorly worded items and checking the performance of the item keys and related distractors. We hoped that editing the tests in this way would help mediate the problem of item redundancy we mentioned above. We also wrote new items which we felt would do a better job at spreading out the population distribution and showing us clearer divisions which could be used for class placement. We piloted our new and modified items by giving them to approximately 100 students from each of our General English program's five levels and then analyzing how well the items performed in this sample. Due to practical concerns, we did not collect this data under uniform testing conditions, but rather asked our colleagues to distribute and collect pilot test papers and mark sheets in their individual classes. In retrospect, it might have been better to give our selected sample of students the test all at one time in more uniform testing conditions, but logistically it was less disruptive to our colleagues to follow the procedures we have described here. We did however organize the timing of item piloting so that we and our colleagues could give the pilot test at the same point in the school term.

In summary, we followed Wilson's framework for constructing measures by generating test items, piloting them with a representative sample population, and analyzing how well the items performed. But we did all of this work without addressing the fundamental need to define and build a construct map. Ultimately, this missing "building block" hindered our ability to create more narrowly targeted test items that might potentially have provided us with more informative data about our students' relative proficiency.

Making it Work

The following section describes in detail the steps taken to improve the reading subtests of the inherited tests. We had to write new test passages to replace unsuitable ones; check that they conformed to predetermined criteria such as vocabulary profile, length, and readability; target items to cover a range of ability levels; pilot the new passages with targeted samples; analyze item and distractor performance from the pilots; and then finally introduce the new passages to the actual tests. This is still an ongoing process, as it should be with test writing, and as the authors are still very much assimilating new ideas and essentially learning new things as we go; there are bound to have been some mistakes made and we welcome any comments and advice that readers may have on our processes and errors.

Schema Issues

As described previously, some of the texts that were being used in the reading subtests when we started on our mission to improve them had a number of problems that made them unsuitable for use, and perhaps the most obvious issue was schema or background knowledge. To illustrate the point, a passage titled *Sacagawea* was being used (see Appendix A), although it is fairly likely that students starting tertiary education in Japan are unlikely to be familiar with Native American history and would be unable to imagine life in very early 19th century America. This issue could potentially mean that even though students can understand the grammar and vocabulary of the passage, they fail to comprehend what the passage is about. And this situation is undesirable for a test that is supposed to be assessing English ability, at least at the placement stage of the testing process.

We soon realized that the only way around passages with schematic problems was to write new passages to replace them. All of the committee members attempted to write new passages,

and although we still made mistakes initially, after a few tries some passages that we felt would not cause problems with schematic understanding were produced. The Earthquake passage (Appendix B) is an example of one of the passages that we produced to replace an inherited passage. We felt that the topic of earthquakes was one that students in Japan would possess the schema to understand without having to use too much of the available time to think about. It should be noted that the passage was written, piloted, and used before the earthquake and resulting tsunami that occurred in northeastern Japan on March 11th, 2011. In retrospect, this topic might also be problematic as it has the potential to upset test takers who have been affected personally by Japan's many deadly earthquakes of recent years. Adding to the existing stress of taking a test is something that should be avoided by test writers.

Length of Passages

At the beginning of the process we did not consider the length of the passages to be an issue with the tests. There were no predefined standard lengths for the passages, but there were some physical constraints dictated by the size of the paper being used. After some thought though, we realized that the length of the passages used needed to be roughly equal across the three tests if comparisons were to be made in order to assess achievement. We therefore made the length for the reading passages roughly uniform across the tests, setting a target length of 250 words.

Vocabulary and Readability

The Sacagawea passage mentioned before had more problems than just schema, not least of which was its vocabulary profile. Tom Cobb's *Vocabulary Profiler v.3* (Cobb, 2006) provides an easy means of assessing the vocabulary profile of a text. It is an online application available on the *Compleat Lexical Tutor* website (Cobb, n.d.) that provides a detailed vocabulary profile for a text, providing information about percentages of words that are in the first 1,000 words of English; the second 1,000 words of English; the Academic Word List (Coxhead, 2000); and off-list words (those that are not in the first three categories).

The profile for the Sacagawea passage (presented in Table 2.) showed that the passage contained a high percentage of off-list words, and further analysis showed that many of these words were proper nouns. In many cases proper nouns are treated as known words because they can be guessed from context (Nation, 2006; Webb & Rodgers, 2009), but in this case they made up a high proportion of the text, with many of them being French in origin. The overall impression even from a native speaker unfamiliar with the schema was one of confusion.

In the Earthquake passage there are fewer off-list words (13.5%) and more words from the first 1,000 words of English, making the passage more comprehensible for the test takers by increasing their predicted coverage of the text. In hindsight it may have been more appropriate to have set goals for all the passages in the tests, however vocabulary is by no means the only factor affecting passage difficulty, and the wordlists used do not completely correlate with student vocabulary knowledge. We used a combination of this information and our own insight into our students' English ability. Overall, the Earthquakes passage appears much easier to understand than passages it was written to replace, not only because of coverage and readability, but also due to the aforementioned schema being much more familiar to the readers.

Table 2. Vocabulary Profiles for Reading Passages

	Sacagawea	Earthquakes
K1 Tokens	75.72%	79.75%
K2 Tokens	5.80%	4.64%
AWL Tokens	1.45%	2.11%
Off-List Tokens	17.03%	13.50%
Total Tokens	276	237

Targeting Item Difficulty

The reading passages that we had written were considered to be ready, so the next step was to write test items for each passage. We planned to pilot the passages, and to do this efficiently and avoid the need for a second round of piloting, we decided it was prudent to write more items than we planned to use in the actual test so that we could exclude undesirable items in the analysis stage. We also decided to write more distractors than would be used in the actual test so as to be able to eliminate poorly performing ones during the analysis stage. Although arguably, items piloted and revised in this way inevitably produce statistically different results in their revised forms, we felt at the time that this process would help us to get better performing versions of the items on our test forms. We could then revisit our changed items in our posttest analyses to determine if in fact the items performed better. Perhaps this approach may have been naïve, but we felt that it helped us understand the item writing process better and called our attention to cases where poorly worded or confusing distractors and keys were problematic.

The items used in the inherited passages tended to cluster together in terms of difficulty, and as this is disadvantageous we purposely set out to write items of a different range of difficulties: some that targeted the high-ability students, some the middle-ability students, and some the lower-ability students. The purpose of doing is was to increase the separation of the students into differing ability groups, which helps with placement on the initial test.

Piloting

We had used all of the instinct that we possessed as teachers and test writers to create as close to perfect passages and accompanying items in the writing stage of the process as we could. We did our best to follow good test writing practices as recommended to us in our background reading. For example, we tried to keep our question stems grammatically simple and short in length; we avoided item choices such as "*none of the above*" or "*both a and d*"; we tried to eliminate ambiguous language; and items with no clear key. We also took steps to format our items uniformly with distractors and keys arranged in order of length and tried to pay attention to the distribution of correct answer patterning among our multiple choice item keys from A to D. However, the proof of our instinct would be strictly gauged by the piloting and subsequent analysis of the obtained data. The piloting, for it to be effective, needed to assess a targeted sample of students from a range of faculties within the university, and a range of levels within those faculties. There were, however a number of constraints on who we could pilot passages with, but we managed to select a suitably stratified sample, with large enough sample sizes at each level to make Rasch Analysis possible. Admittedly, it might have been better to insert pilot items in the test forms themselves to see how they might perform under real testing conditions. But, we felt, in our context, that it would be better to move more quickly to address items that seemed problematic both to us and to our students.

Analysis and Interpretation

We used the WINSTEPS software application to conduct Rasch analysis on our test and pilot items (Linacre, 2011). We first examined the test Wright map to look at the item difficulty values and to check if our targeted pilot items performed as expected. We paid attention to both the relative distance between the items and their alignment to the person ability measures. We tried to select items that helped separate our students into distinct ability groupings. Next, we examined how the item distractors performed and looked for unexpected patterns in the relationship between the percentage of students choosing certain distractors and their correspondent person ability estimates. Unexpectedly attractive or less attractive choices were examined more closely to see if there were design problems in either the item prompt, stem, key, or distractors. We also examined the item fit statistics which are indicators of how well our data matched up with the Rasch measurement model. We looked first at the mean square infit statistic, a measure that is "inlier" sensitive, meaning that the value reflects changes when item difficulty and person ability are more closely aligned. In contrast, the outfit value is more sensitive when the distance between person ability and item difficulty is larger. In our context, we paid more attention to the infit statistic when pilot items were embedded in a full test form, reasoning that removing outliers from the analysis would produce a clearer indication of item viability. Not unexpectedly, we have seen many cases of low-level students' responses that indicate they are guessing or that they have given up and failed to select any answer at all. When we piloted items with smaller groups we looked more at the outfit statistic reasoning that outliers influential enough to cause misfit in a smaller pilot test sample size should be examined more closely. Next, we looked at two item discrimination statistics, both the discrimination value produced by WINSTEPS and the more traditional point biserial correlation values of the key and distractors. In both cases item discrimination values indicate how well a correct answer on a single item corresponds to a better score on the entire test. Based on the cumulative data, we decided which distractors to discard and which items would be promising candidates to put into the test. And as a final step, we graphed the item characteristic curves for each potential new item to check that the items worked well for across all student ability levels.

Student Perspectives

One important step we took to improve our items was to ask our students for qualitative feedback on the tests. We did so not because we lacked confidence in our instincts as teachers and test makers or because we were skeptical of the statistical data provided by our Rasch analysis software. Instead, we recognized the need to triangulate our data sources and hoped that the qualitative data would shed some light on how our students were processing our test items. In retrospect, this decision was fortuitous as even with a limited number of informant responses, we were able to gain new insights into how students at different proficiency levels perceived and processed our test items.

The data we collected during the piloting phase really helped us to see why certain items performed the way they did. We collected this qualitative data by interviewing a cross section of students and eliciting written comments from them immediately after they completed working with the pilot passages. Here, we will share some of the written entries related to the reading passages for our newly piloted items and will tie them back to our statistical data.

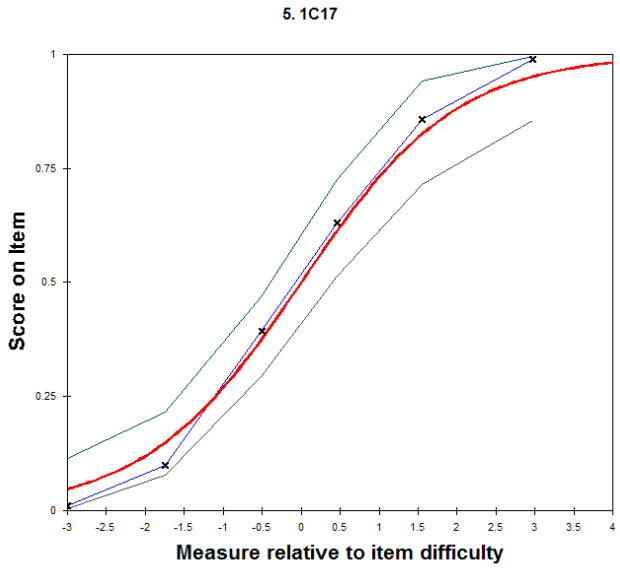

Figure 3. Item characteristic curve of memo passage item #11 (1C17).

Figure 3 above shows the item characteristic curve for item number 11 from one of our piloted reading passages called "Memo." The pilot form of this reading passage and its associated items can be found in Appendix C of this paper. The blue segmented line in this graph represents our students' actual performance on this item relative to item difficulty. The horizontal axis represents the person ability measures on an ordinal scale and the vertical axis represents the item score. This line is sandwiched between two lines which represent the upper and lower bound 95% confidence intervals. When the item characteristic curve falls inside these boundaries, it indicates that the actual performance of the test item conforms to the Rasch measurement model. The key idea here is that the more the data fit the model, the more the assumptions about how the model functions can be trusted and the more confident we can be about the quality of measurement of our instruments.

The smooth red line represents the model itself. Lower-ability students are on the left side of the curve and higher-ability students are on the right side with the zero value representing the person mean. We can see the probabilistic aspect of the model best in this type of graph as students of lower ability are less than 50% likely to correctly answer items that are more difficult than their measured ability, whereas higher-ability students have a higher than 50% likelihood of answering the same item when it is easier than their estimated ability.

Returning now to the item performance in Figure 3, the graph indicates that the item conforms fairly closely to the model. From a statistical point of view this item seems to be working well. However, our qualitative student feedback data paint a very different picture and reveal some unexpected insight into the thought processes of our test takers. Let us compare the comments made by R, a student in the highest-level course in our program with those of T, a lower-level student directly after taking the pilot items.

> *As I skimmed, I tried to understand who is involved, what they are doing and why. In the end, I got the idea from the Message section.* (Student R, a high-level student; answer = correct)

> *I had no idea so I looked for the same words in the passage* (Student T, a low-level student; answer = incorrect)

It appears that Student R had a clear strategy attempting to imagine who, what and why; whereas student T was not able to understand the gist of the passage and chose a distractor which

contained an identical word that she could pick out from the passage prompt. So in this case the higher-level student took a more sophisticated reading approach and chose the correct answer.

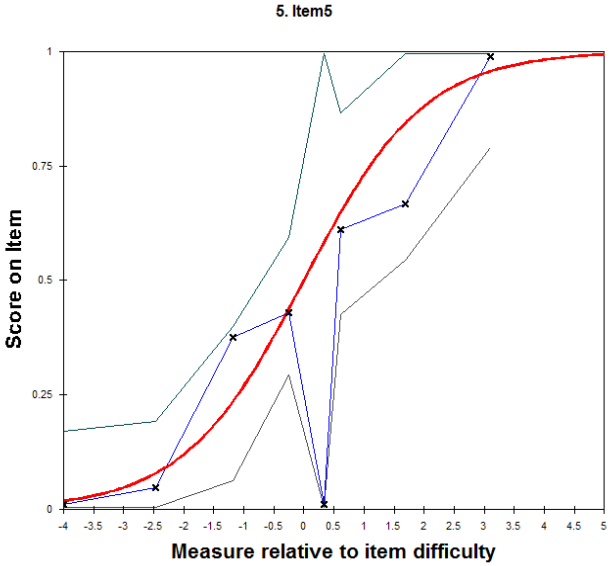

Figure 4. Item characteristic curve of Earthquakes passage Item #5.

Figure 4 above shows us a much messier picture and a somewhat unexpected result. The reading passage and piloted items can be found in Appendix B of this paper. In this item characteristic curve, we can see that the lower-level students indicated by the left half of the blue line are performing beyond the ability level predicted by the Rasch model. In contrast, the higher-ability students are doing the exact opposite. They have actually underperformed and missed this item when their person ability measures indicate that they should probably have gotten this one correct. Once again a look at some student comments sheds some light on why this might have occurred.

> *I kept reading the sentences around the part in question. In the end, I had to make a guess.* (student R, a high-level student; answer = incorrect)

> *I saw "deep". That's why I chose "far beneath the surface".* (student S, a low-level student; answer = correct)

As she mentions in her comment, high-level student R, the same student who was able to answer the Figure 3 item correctly, seemed to have spent a great deal more time trying to answer this question than the lower-level student S. However, student S chose the correct answer even though she reported that she did not really read the passage carefully. Additional student comments like these caught our attention and made us reconsider the legitimacy of including this question in future revisions of this test.

Indeed we were often surprised by how differently our students approached test items and how often their comments helped us understand new ways of looking at the questions themselves. An amusing example of this phenomenon can be found in one student's reaction to item 9 of the Memo passage (Appendix C). The question stem reads, "Who is Thomas Read?" The item writers expected the students to scan the hotel memo passage and select *(A) a hotel guest* after noticing the name, Mr. Thomas Read at the top center portion of the hotel memo prompt. Instead Student Y wrote down this revealing comment.

This is so unrealistic! Why doesn't the person use cell phone to communicate? (student Y; answer = incorrect)

In fact, student Y got all the other questions correct except for this one. She seemed to have understood all the information on the memo but did not understand how a hotel memo works. For students who grew up with cell phones, a hotel memo is extremely unfamiliar. Y's comment brings us back to the need for clearer test specification and detailed construct mapping at the start of the test development process. Is this test made to evaluate students' cultural knowledge or English proficiency? It seems to us that this would be a more fruitful discussion in the initial stages of test development.

What We Achieved

It's clear that there is still much more work to do to improve the quality of our tests, but looking back we can point to positive indicators of progress. Our piloting and revision work with the reading passages has helped us to get better person separation in our tests which should greatly help us with placement decisions. Our efforts to target item difficulty to different student ability levels has led to better alignment of our item difficulty and person ability means and fewer outliers in the data who fall outside the measurement range of the test items. More importantly, although there continue to be a large number of items clustered about the person mean, the item difficulties now align much better with gaps in person ability levels. This is a promising result since it indicates that the test items are doing a much better job of placing moderate ability students into different groups.

Conclusion: Future Directions

Despite some tangible progress in our test development project, it is clear to us that there is still much work to be done before our tests become more trustworthy. Although equating work was initially a top priority in this project, item revision was undertaken with the goal of achieving better model fit. We would like to argue that this project might best move forward by taking a step back. In short, we would propose returning to the first of Wilson's building blocks and call for a more concerted effort to create a detailed construct map. The logic here is simple. By breaking down language proficiency into smaller subskills, more precisely targeted items should be easier to write which in all likelihood would produce clearer statistical results. Additionally, a more detailed construct map could help reinforce objectives in the teaching curricula. Finally, we would also like to wholeheartedly endorse the practice of eliciting qualitative data from students. As we have suggested in this paper, students can be an invaluable source of insight into how test items function and can help guide is in both our item writing and revision processes.

We would like to conclude on a positive note. Although we were very much neophyte test makers at the start of this testing project, we were able to move the project forward and feel that our current tests, though far from complete, are in better shape than when we began. We were able to work cooperatively with the support of senior faculty and implement systematic steps to analyze and revise significant portions of these tests. Finally, we would be remiss if we failed to mention our own professional development. With effort and persistence we have become more conversant in the language and theory of language assessment and hope that we have shown that teachers can, in fact, play a leading role in large-scale assessment work.

Bio Data

Jay Wong is currently a lecturer at Ritsumeikan University. His areas of interest include second language writing fluency development, motivation, and language testing. <jglwong68@gmail.com>

Andrew Atkins is a lecturer at Kinki University, Japan. He is a doctoral candidate at Temple University and his dissertation research is focused on reading fluency development. Other research interests include international education, CALL, research methodology and language testing. <andrew@kindai.ac.jp>

Taka Okazaki is a lecturer at Kinki University. His areas of interests include language teaching, language policy, language revitalization and indigenous education. <oktaka@kindai.ac.jp>

References

Bond, T., & Fox, C. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Cobb, T. (n. d.). Compleat Lexical Tutor. [http://www.lextutor.ca/].

Cobb, T. (2006). Web Vocabprofile [Version 3.0. Accessed from http://www.lextutor.ca/vp/].

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Fulcher G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Abingdon, U.K.: Routledge.

Linacre, J. M. (2011). *Winsteps® Rasch measurement computer program.* Beaverton, OR: Winsteps.com

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59–82.

Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning, 59*(2), 335-366.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Appendix A

Sacagawea Reading Passage

How did a young Native American woman named Sacagawea become a guide and interpreter for the famous explorers Lewis and Clark? In 1801, when she was thirteen, Sacagawea was captured by a neighboring tribe who later sold her to a French-Indian trader named Toussaint Charbonneau. In 1804, President Jefferson ordered Merriweather Lewis and Captain William Clark to form an expedition to search for a route to the Pacific Ocean. The expedition needed an experienced interpreter and eventually the leaders hired Charbonneau to talk with the native people. When they found out that he had a native American "wife," they asked him to bring her along. At the time, Sacagawea was sixteen and pregnant.

Sacagawea proved to be very important to the expedition's success. She guided the expedition to her native village and, because she could speak the native language, Lewis and Clark were able to buy horses from the people there and ask them to supply a guide to lead the expedition over the rugged Bitterroot Mountains. Lewis praised Sacagawea's steady thinking, and gave her credit for saving the expedition's research materials when one of the boats overturned in the Missouri River. Along the way, she gave birth to a son and carried him the entire journey on her back. Because it traveled with a woman and a child, the Native American tribes the expedition met quickly understood that the expedition was peaceful.

Sacagawea received no payment for the work she did for the expedition, but she has been rewarded by history. A statue of Sacagawea stands in the Capitol Building in Washington, D.C., and her face is on the back of the U.S. silver dollar coin.

[5]下線部 (a) form an expedition a. send a group on a mission b. receive an order from a nation's leader c. put together a group for a long journey d. make an agreement with the government	[7]下線部 (c) it a. the boat b. the child c. Sacagawea d. the expedition
[6]下線部 (b) them a. the guides b. the villagers c. Lewis and Clark d. the expedition's members	[8]下線部 (d) been rewarded by history a. received money b. become famous c. stood in the Capitol Building d. been taken to Washington, D.C.

Appendix B

Earthquakes Reading Passage

On March 26, 2000, the Kingdome in Seattle, Washington was torn down. The structure was (a) flattened to make room for a new sports stadium. The blast from the falling Kingdome caused the Earth to shake as if an earthquake had happened. Scientists placed more than 200 seismic recorders in the Earth to measure the movement. They found which parts of the city shook the most. This information helped them (b) predict which parts of the city would be damaged in a real earthquake. On February 28, 2001, (c) the real thing happened. The Nisqually earthquake was 6.8 on the Richter scale. It was a (d) deep focus tremor that damaged the same parts of Seattle that scientists had predicted from the Kingdome blast. (e) Such earthquakes start deep in the Earth and often cause few aftershocks. In fact, the Nisqually quake started 37 miles below the surface on the Juan de Fuca plate and had only four. Another earthquake in California that was close to the surface had over one hundred twenty aftershocks. Scientists do not know why the deep earthquakes have fewer aftershocks. Seismologists plan to set off explosives in the ground near Seattle aimed at the slab (plate). The shockwaves from the blast will bounce off the slab and give (f) them an idea of where the plate is and how it is moving. This will give them more information in case another real earthquake hits the area.

1.下線部 (a) flattened e. emptied f. moved g. destroyed h. closed	2. 下線部 (b) predict a. arrange b. record c. choose d. see
3. 下線部 (c) the real thing a. a blast b. a measurement c. an aftershock d. an earthquake	4. 下線部 (d) tremor a. earthquake b. blast c. prediction d. measurement
5. 下線部 (e) Such a. close to the Earth's surface b. far beneath the Earth's surface c. above the Earth's surface d. on the Earth's surface	6. 下線部 (f) them a. seismologists b. explosives c. shockwaves d. blasts

Appendix C

Memo Reading Passage

MEMO				
To:	Mr. Thomas Reed			URGENT ☑
Room Number:	1205			
Date:	11/14	Time:	10:15	A.M. ☑ P.M. ☐
WHILE YOU WERE OUT				
Mr./Mrs./Ms.	Selina Garcia			
From:	Ellison Computers			
Phone:	408	321-6577	5678	
	Area Code	*Number*	*Extension*	
	TELEPHONED	☑	PLEASE CALL	☐
	CAME TO SEE YOU	☐	WILL CALL AGAIN	☑
	RETURNED YOUR CALL	☐	WANTS TO SEE YOU	☐
Message:	Meeting time with Maxwell has been changed to an hour earlier to 2:30 p.m.			
	You will need to reanalyze the technical data sooner than planned.			
	Can you please update the sales presentation slides by noon?			
	Signed:	Jessica Wilcox, Assistant Manager		
San Francisco Tower Bridge Hotel *For Business or Pleasure: Call 1-800-415-9965 for Reservations*				

[7] Who made the telephone call?

(A) Mr. Reed

(B) Ms. Garcia

(C) Mr. Ellison
(D) Ms. Wilcox
(E) Mr. Maxwell

[8] When was the message taken?
(A) In August
(B) In October
(C) In December
(D) In November
(E) In September

[9] Who is Thomas Reed?
(A) A hotel guest
(B) A telephone caller
(C) A technical reporter
(D) A reservations clerk
(E) An assistant hotel manager

[10] What should Mr. Reed do now?
(A) Meet Ms. Garcia
(B) Send a fax to his office
(C) Meet with Mr. Maxwell
(D) Wait for a telephone call
(E) Write back to Ms. Wilcox

[11] Why was this call made?
(A) To return a call
(B) To update a computer
(C) To make a reservation
(D) To send technical data
(E) To change an appointment