Developing an Academic English Program Placement Test: A Pilot Study

Trevor A. HOLSTER
J. LAKE

福岡女子大学文学部・国際文理学部紀要「文藝と思想」
第76号　pp.23～46 抜刷　2012年2月

福岡女子大学文学部・国際文理学部紀要
「文藝と思想」第76号　2012年2月　23～46頁

Developing an Academic English Program Placement Test: A Pilot Study

Trevor A. HOLSTER
J. LAKE

Abstract

Placement tests play an important part in language programs by matching students to instruction at an appropriate level, but optimally should be developed to closely reflect the objectives and students of specific institutions. When such targeting is not possible, general tests of lexico-grammar can be used as generic placement measures, but matching of test difficulty and person ability is still necessary to ensure that measurement is precise enough to statistically separate groups of students. A fifty-item multiple-choice cloze test was developed and piloted, then analyzed with Rasch measurement to investigate its suitability as a placement instrument for the Fukuoka Women's University Academic English Program. The preliminary analysis indicates that the test will adequately separate high and low groups of students and identify students requiring remediation, but it is recommended that listening, reading, and vocabulary synonymy items also be developed for a longer test battery in order to better evaluate higher proficiency students and provide improved diagnostic feedback.

Introduction and Background

Although language programs employ tests for various reasons (Bachman & Palmer, 2010; Brown, 2005), "the nature and validation of placement tests is rarely

discussed in the language testing literature, yet placement tests are probably one of the commonest forms of tests used to make decisions across the institution rather than individual classes" (Wall, Clapham, & Alderson, 1994, p. 321). Placement tests, however, differ in important ways from the classroom assessments familiar to teachers. Classroom assessments are typically criterion referenced, measuring learning against criteria related to the specific objectives of the class (Brown & Hudson, 2002). "Criterion-referenced tests (CRT) are designed to enable the test user to interpret a test score with reference to a criterion level of ability or domain content" (Bachman, 1990, p. 74). Such tests are often used for diagnostic purposes to determine what is already known and what needs to be learned, administered as pretests to aid in curriculum planning, or midterm to guide remediation. Classroom tests are also used to assess achievement, administered at the end of a course to measure the amount of learning in the class. This may be expressed as a percentage score range from 0%, for no learning, to 100%, indicating that a student has learned everything taught. Ideally, the criterion-referenced pretest score should be very low and the achievement test score very high, showing effective learning of course content (Brown, 2005).

In contrast, placement and admissions tests are usually norm-referenced tests (NRT), intended to spread students out along a normal distribution curve to magnify differences among students for purposes of grouping or comparison. Such tests measure broad abilities such as aptitude, overall proficiency, or general language skills, rather than being based on the content of a particular class. Table 1 compares six different characteristics of norm-referenced and criterion referenced tests.

Placement tests aim to sort students into homogeneous groups for instruction so that they can benefit from classes that "are neither too difficult nor which involve wasteful duplication of earlier-learned content" (Chauncey & Frederiksen, 1951, p. 109), allowing teachers to "focus on the problems and learning points appropriate for that level of student" (Brown, 1996, p. 11), making the level of learning more appropriate for students and teaching more efficient for teachers. Although commercially produced placement tests promise to save considerable time relative to in-house development, Westrick investigated one of these, finding

Table 1 Differences between Norm Referenced Tests (NRT) and Criterion Referenced Tests (CRT) (adopted from Brown, 2005)

Characteristic	NRT	CRT
Type of interpretation	Relative (percentile)	Absolute (percentage)
Type of measurement	To measure general language abilities or proficiencies	To measure specific objectives based language points
Purpose of testing	Spread students out along a continuum of general abilities of proficiencies	Assess the amount of material known, or learned by each student
Distribution of scores	Normal distribution of scores around a mean	Varies, usually non-normal (students who know all of the material should all score 100%)
Test structure	A few relatively long subtests with a variety of question contents	A series of short, well defined subtests with similar question content
Knowledge of questions	Students have little or no idea what content to expect in questions	Students know exactly what content to expect in test questions

that:

> the test results failed to discriminate among students of varying proficiencies. Narrow ranges, low score reliability estimates, and large standard errors of measurement characterized the results. Item analysis revealed that most of the test items did little to separate high and low scoring students. (Westrick, 2005, p. 71)

Westrick therefore argued for in-house placement test development based on existing curriculums to "make the assignment of students to specific classes more logical and defensible" (Westrick, 2005, p. 90). One benefit is that while commercial tests must measure the full range of ability found across all institutions, an in-house test need only measure the ability range within a specific institution. Items outside this ability range will not be measuring anything, creating a less efficient and reliable test:

> A placement test must be more specifically related to a given program, particularly in terms of the relatively narrow range of abilities assessed and the content of the curriculum, so that it efficiently separates the students in level groupings within that program. (Brown, 1996, pp. 11-12)

By definition, commercial tests cannot be closely matched to any specific institution. "No one placement test will work for every institution, and the initial assumption about any test that is commercially available must be that it will not work well" (Hughes, 2003, p. 16). Brown (1995, 1996) agreed that placement tests should be closely integrated with the teaching curriculum, that is, fit with the goals and objectives of the language program and what is happening within the classrooms. Thus, the test's psychometric properties must support the spreading out of students, while the content of the items should reflect the content of the curriculum at various points and be sensitive to student learning through the levels (Brown, 1989). Since placement tests aim to spread students into different ability levels, Brown and Hudson (2002) argued that "placement decisions are best made using norm-referenced tests," creating tension between the requirements of content validity and psychometric validity. They noted, however,

> We make this statement in light of the present state of knowledge about how language is learned. However, if one day we come to understand clear-cut stages throughout the language learning process, and/or if a truly hierarchical curriculum is developed some day, there is no theoretical reason why we could not have a CRT placement test.

Rasch measurement (Bachman, 1990; Bond & Fox, 2007; Szabo, 2008) provides a theoretical basis to "reconcile norm-referenced and criterion-referenced approaches to assessment" (McNamara, 1996, p. 198) while Clark noted the complexities involved and states that "the Rasch model is ideally suited for measurement for placement purposes" (2004, p. 66). Rasch analysis is a probabilistic model derived from the simple conceptual insight that greater person ability increases the probability of a correct response, while greater item difficulty decreases the probability of a correct response, but that real-world data do not follow a strictly deterministic pattern of success and failure. From these fundamental premises, very sophisticated tools can be derived to analyze individuals' responses to items, individual items' contribution of information to the whole test, and how the information is organized throughout the test. Rather than relying on a single raw-score summary for each person, the information contributed by each person's response

to each item can be considered.

From the Rasch perspective, the crucial consideration is the precision of the test at important points of difficulty for the construct being measured and the precision in determining a learner's ability. Increased precision requires increased information at the ability level of interest, i.e. near the cut points in a placement or admissions test (Embretson & Reise, 2000; Thissen & Wainer, 2001). Rasch analysis thus provides a variety of statistics for persons, items, and the test as a whole (Bachman, 1990; Bond & Fox, 2007; Embretson & Hershberger, 1999; Embretson & Reise, 2000). Thus, in-house tests and Rasch analysis will allow feedback about learning and teaching in the academic English program (AEP) far beyond what is possible with classical analysis of raw scores and improve the tailoring of tests to learner needs and abilities.

Placement test development ideally begins with needs analysis and pilot testing to match the test to the institutional objectives and ensure that what is tested relates to what is taught, but such clearly defined curriculum specifications are not typical of Japanese universities, as detailed by Inoue (2006). Delays in developing a detailed curriculum at FWU thus complicate the development of placement tests, making tests of general lexico-grammar attractive because such tests can address universal cognitive traits that are applicable to any curriculum (Pienemann, 1984; Pienemann, Brindley, & Johnston, 1988), while addressing the issues of practicality and reliability raised by Bachman and Palmer (1996). As words are a fundamental building block of language, vocabulary is a strong predictor of proficiency and a useful indicator for placement (Meara & Jones, 1988). As expressed by Stahl (1999, p. 3), "One of the oldest findings in educational research is the strong relationship between vocabulary knowledge and reading comprehension," but vocabulary is now seen as important for the development of all language skills, not just reading. "Vocabulary is an essential building block of language," noted Schmitt, Schmitt, & Clapham, (2001, p. 55). The same point was made by Read (2000, p. 1), "Words are the basic building blocks of language, the units of meaning from which larger structures such as sentences, paragraphs and whole texts are formed." The importance of vocabulary for ordinary communication is known intuitively by most learners; Krashen (1989, p. 440) stated, "A large

vocabulary is, of course, essential for mastery of a language. Second language acquirers know this; they carry dictionaries with them, not grammar books, and regularly report that lack of vocabulary is a major problem."

In addition to aiding placement decisions, vocabulary measures can guide vocabulary instruction, identify appropriate graded readers for extensive reading and inform material and textbook selection. Therefore, a vocabulary subsection was considered a priority in order to provide estimates of vocabulary knowledge at different student levels. However, much recent scholarship emphasizes the closely intertwined relationship between grammar and vocabulary, and the patterned and formulaic nature of language (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Howarth, 1998; Hunston & Francis, 2000; Mukherjee, 2001; Sinclair, 2004; Willis, 2003; Wray, 2000, 2002; Wray & Perkins, 2000). There has also been an explosion in corpus-based research and the use of corpora in language teaching, as the title of Sinclair's (2004) edited collection suggests. This expansion in research and literature highlights the centrality of vocabulary and its relation to other language issues and relationships. Hunt and Beglar (2005, p. 24) noted, "As teachers and researchers have come to understand the role of the lexicon in language learning and communication, the increased attention to vocabulary teaching has become more important."

One ubiquitous method of testing lexico-grammar is the cloze procedure, where a word is deleted from a sentence or longer passage. Cloze tests are simple to construct and correlate strongly with general language proficiency, so they were for a time considered "almost as a language testing panacea" (Hughes, 2003, p. 189). The theoretical basis of the cloze procedure originated in gestalt psychology (Taylor, 1953), which provides the insight that people can cognitively understand a greater "whole", even when many of the "parts" are missing. Thus, the gestalt concept of "closure" drives readers to fill in missing gaps in text. The cloze procedure is thus an integrative item type, where knowledge of vocabulary, grammar, and pragmatics all contribute to responses rather than aiming to measure knowledge of word meaning in isolation. Higher proficiency learners can do this better than lower proficiency learners and tend to make judgments based on more context than lower ability learners who make judgments based on more limited context

(Brown, 2002). Jonz (1976) developed the multiple-choice (M-C) cloze, and found that they could be scored quickly and reliably.

Holster and DeLint's (in press) Longman Vocabulary Level test (LVL) provided a potential source of M-C cloze test items with an extensive database of concurrent and longitudinal results, using Rasch analysis to validate and equate the tests (Bond & Fox, 2007; McNamara, 1996). However, this test was originally developed as a research instrument to measure the vocabulary gains of novice level learners, so considerable revision and pilot testing was undertaken to develop items suitable for a placement test at FWU.

Test Development and Results

One concern with the LVL was its use of a clustered item format, where each test form comprised 10 clusters of 5 items, as illustrated in Figure 1. This is potentially problematic because correctly answering one item changes the odds of successfully answering other items within the cluster, in violation of the assumption of independence underlying many common statistical analyses (Field, 2009, p. 133). The *Winsteps* software package for Rasch analysis (Linacre, 2010) quantifies dependency effects through inter-item correlations, as shown in Table 2, listing the 10 largest inter-item correlations. Although all of the 10 largest dependencies occur within item clusters, the magnitude is small. The largest correlation was only .28, meaning that only about 8% of variance was shared between these items, far too small to substantively affect results.

Although serious dependency problems have been avoided in this instance, the creation of parallel forms is problematic because every test form must be piloted in order to check for dependency effects. Additionally, the cluster format greatly complicates test equating through shared items because combining existing items into new clusters results in changed distracters, potentially destabilizing item difficulty. Development of a more conventional discrete item test was thus necessary.

Use of Holster and DeLint's (in press) items as the starting point for an item bank was supported by the results of Rasch item analysis. As well as a reliability

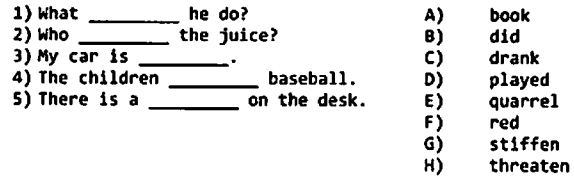

Figure 1 **Clustered item format. In this example, five item stems share eight answer choices, resulting in dependency between items, where knowing the correct answer to on item can improve the odds of a successful response to other items.**

Table 2 Largest Standardized Residual Correlations

Correlation	Item Number	Item Number
.28	82	85
.23	36	39
.23	21	23
.21	97	98
.20	24	25
.19	88	89
.18	83	84
.17	21	22
.17	81	84
.17	61	63

coefficient analogous to the commonly used Cronbach's alpha, Rasch analysis provides detailed analysis on the performance of individual items and persons beyond that possible from classical raw score analysis (Bachman, 1990; Bond & Fox, 2007). Table 3 shows the performance of persons administered the original clustered format tests in 2008 and 2009. Person reliability of .90 indicates stable rank ordering of persons, and the separation index of 2.98 means that the sample of persons can clearly be separated into at least two, and perhaps three, statistically distinct bands of ability.

Rasch measurement also provides measures of data-model fit, indicating how closely the data match the predictions of the psychometric model. Mean-squared (MNSQ) fit statistics have an expected value of 1.00, indicating 100% of the predicted randomness in the data and a range from 0.00 to infinity. The infit statistic is information weighted, emphasizing responses where person ability and item difficulty are closely matched, providing a crucial indicator of the quality of measurement, while the outfit statistic is unweighted, providing an indication of

inconsistency in outlying responses. From Table 3 we can see that the mean outfit statistic of 1.06 and standard deviation of 0.68 indicate some unpredictability among outlying responses, but the infit mean-squared value, at 0.99, is extremely close to the expected value and the standard deviation of 0.19 indicates that effective measurement of persons is possible.

Table 4 shows concurrent evidence for construct validity, comparing measures from the LVL, the Vocabulary Levels Test (Beglar & Hunt, 1999; Schmitt, et al., 2001), listening and reading scores from the TOEIC Bridge (ETS-IIBC, 2002), and word translation from Japanese to English and English to Japanese. The LVL

Table 3 Summary of 2557 Measured Persons

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
M	29.0	50.0	55.55	3.74	0.99	0.0	1.06	0.1
SD	9.2	0.0	12.37	0.70	0.19	1.1	0.68	1.1
Max.	49.0	50.0	98.95	10.32	1.82	4.5	9.90	6.4
Min.	4.0	50.0	16.61	3.35	0.52	–3.3	0.16	–2.2

Real RMSE 3.93 True *SD* 11.72 Separation 2.98 Person Reliability .90
Model RMSE 3.81 True *SD* 11.77 Separation 3.09 Person Reliability .91
SE Of Person Mean = 0.24

Table 4 Correlations between Concurrent Tests

		Vocabulary Levels Test	TOEIC Bridge Listening	TOEIC Bridge Reading	Japanese-English Translation	English-Japanese Translation
Longman Vocabulary Level	Corr.	.718	.573	.625	.742	.654
	Sig.	.000	.000	.000	.000	.000
	N	105	123	123	119	119
Vocabulary Levels Test	Corr.		.537	.603	.753	.700
	Sig.		.000	.000	.000	.000
	N		175	175	198	198
TOEIC Bridge Listening	Corr.			.577	.562	.546
	Sig.			.000	.000	.000
	N			207	191	191
TOEIC Bridge Reading	Corr.				.573	.554
	Sig.				.000	.000
	N				191	191
Japanese-English Translation	Corr.					.783
	Sig.					.000
	N					238

measures correlated most highly with Japanese to English translation, with shared variance exceeding 55%. The correlation with the reading section of TOEIC Bridge was moderate to strong, with approximately 40% shared variance. Attenuation due to measurement error means that the true correlations will be even higher, but the raw correlations reported provide supporting evidence that the LVL provides a pool of lexico-grammar items suitable for placement testing.

However, the sample used in validation of the LVL included a considerable number of very low proficiency learners, meaning that the easiest items on the LVL would contribute little to measurement of the more able FWU population. The 60 most difficult item stems were therefore rewritten as discrete items, as shown in Figure 2. These were piloted in semester 2 of 2010 and the best performing 50 items were selected. Many-faceted Rasch measurement (MFRM) (Linacre, 1994) was used to compare clustered and discrete formats (Lake & Holster, 2011).

In conventional 2-faceted Rasch measurement, the odds of success are modeled as a function of the difference between person ability and item difficulty, so that greater person ability increases the probability of success and greater item difficulty decreases the probability of success. Persons have 50% odds of success when person ability and item difficulty are equal, with odds approaching 0% on extremely difficult items and approaching 100% on extremely easy items. However, MFRM allows other facets to be measured. In the initial analysis, a 5-faceted model was specified, with the probability of success modeled as the relationship between person ability, time of administration, test version, order of administration, and item difficulty. Figure 3 shows the facets measurement rulers. The logit measurement scale on the left shows person ability ranging from -1.6 logits to 2.1 logits. A person higher on the scale has a higher probability of a successful response, indicated by the "+Person" column label. The other facets are negatively oriented, so the likelihood of a successful response decreases higher on

```
1. I thought he was going to _________ me.
A) hit        B) inquire    C) advice    D) borrow
```

Figure 2 Revised discrete item format. The revised item format eliminates inter-item dependency because distracters are no longer shared between items.

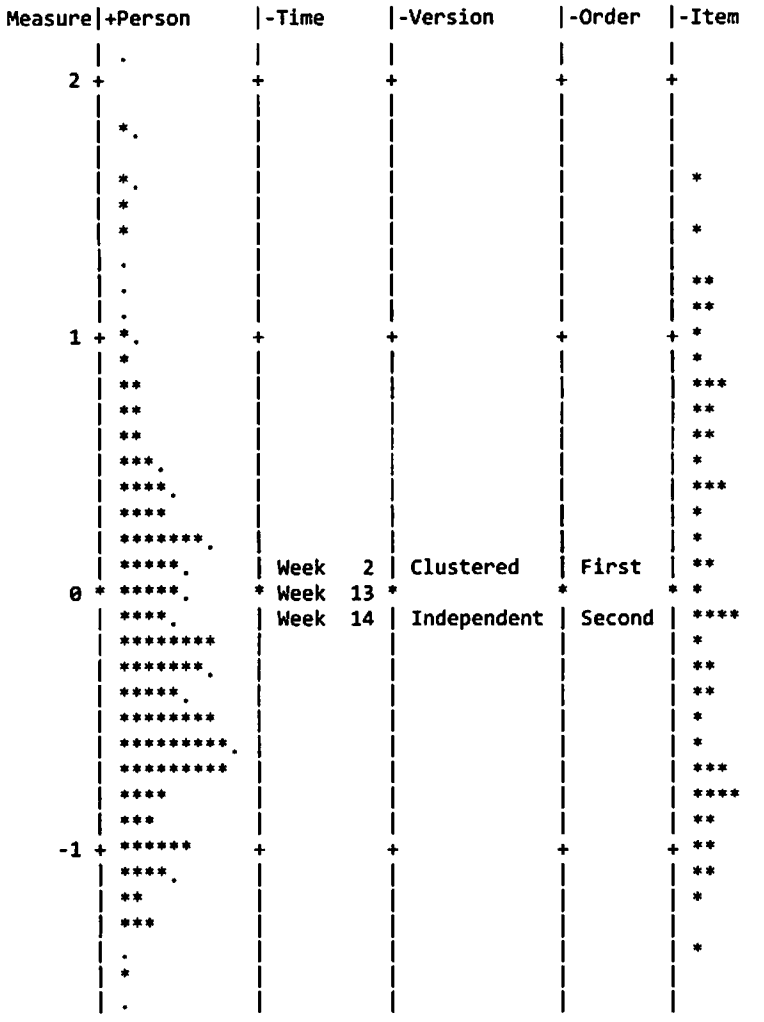

Figure 3 **Facets measurement rulers. The probability of success was modeled as a function of person ability, time of administration, test version, order of test administration, and item difficulty. These five facets are mapped on a common measurement scale.**

the scale, indicating greater item difficulty. Thus, items became relatively easier in week 14 than week 2, and the second administration was relatively easier than the first administration, evidence of both a practice effect and learning. Of central concern was the effect on item difficulty of the change to independent items, and Figure 3 shows the revised format was somewhat easier than the clustered format, raising concern over finding enough items of sufficient difficulty for the FWU population.

Investigation of interaction between items and test format raised a further issue, illustrated by Figure 4. Although, on average, the cluster format items were about 0.30 logits more difficult, this varied considerably, with some discrete items more difficult than their clustered counterparts and some clustered items more than 1.0 logit greater in difficulty. This instability between the two formats raises questions about the contributions of stem, target word, and distracters to item difficulty. The clustered format complicates this, so this evidence supports the change to the simpler discrete item format.

To increase the size of the item pool and allow further investigation of the contributions of stems and distracters to item difficulty, two further 50 item forms were created. In both revised forms, students were presented with the same four answer choices as version 1, but version 2 tested a different target word from

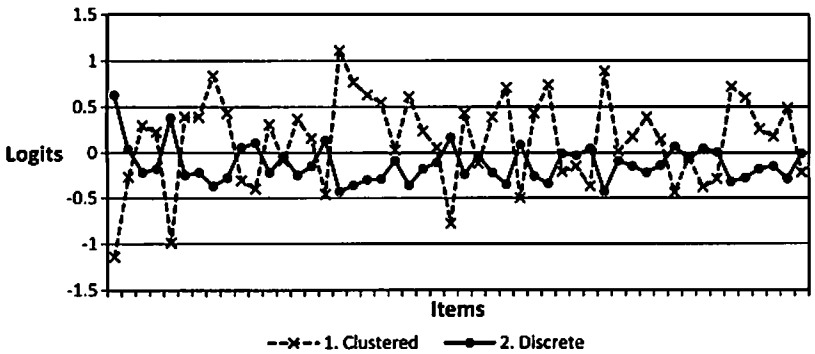

Figure 4 Bias/interaction report comparing clustered versus discrete item difficulties. Although the discrete item format was, on average, easier than the clustered version, relative item difficulties were not stable between the two formats.

among the four answer choices used in version 1, while version 3 tested the same target word using a different gapped sentence stem. Linking of the three versions was achieved by creating sub-forms comprising items 1 to 20 from one version and items 21 to 50 from a different version, giving nine sub-forms in total. Analysis was conducted as a 2-faceted model using *Winsteps* (Linacre, 2010). The 2010 data had been scaled to approximate results from the vocabulary levels test (VLT) (Schmitt, et al., 2001), providing a user-friendly scaling, so this scaling was maintained.

However, no students failed on four of the new items, requiring their removal due to mismatch to the target population. Figure 5 shows the measurement rulers for the remaining 146 items and 457 persons. Although the mean item difficulty is 1500, the mean person ability is close to 1800, and there are relatively few items above the 2500 level, so many of the easier items can be removed without compromising the quality of measurement.

Table 5 shows summary statistics of the 455 measured persons, with person reliability of .86 indicating acceptable rank ordering of persons for a 50 item test, and infit and outfit mean-squared statistics precisely matching the expected mean value of 1.00. The standard deviation of infit and outfit were 0.13 and 0.36 respectively, suggesting some unexpected outlying responses but consistent with effective measurement.

Table 6 shows summary statistics of the 146 remaining items, with mean infit and outfit values of 1.00 and 0.99 respectively, and standard deviations of 0.07 and 0.16, indicating very good data-model fit for items. Comparison of the fit statistics of items and persons suggests that the items effectively measure a unidimensional latent trait, but that some persons behaved unpredictably. Overall, the items performed well, providing an item pool sufficient to develop an operational form suitable for placement at FWU.

Table 7 shows item statistics for the 50 items with the lowest point-measure correlation, indicating how strongly each item correlated with the overall test. The predictions of the model (PT-MEASURE EXP) are typically greater than .25, but 16 items have reported correlations lower than this, indicating that they are not effectively discriminating between low and high ability persons. However, the 16 weakly performing items are only 11% of the total item pool, only four items have

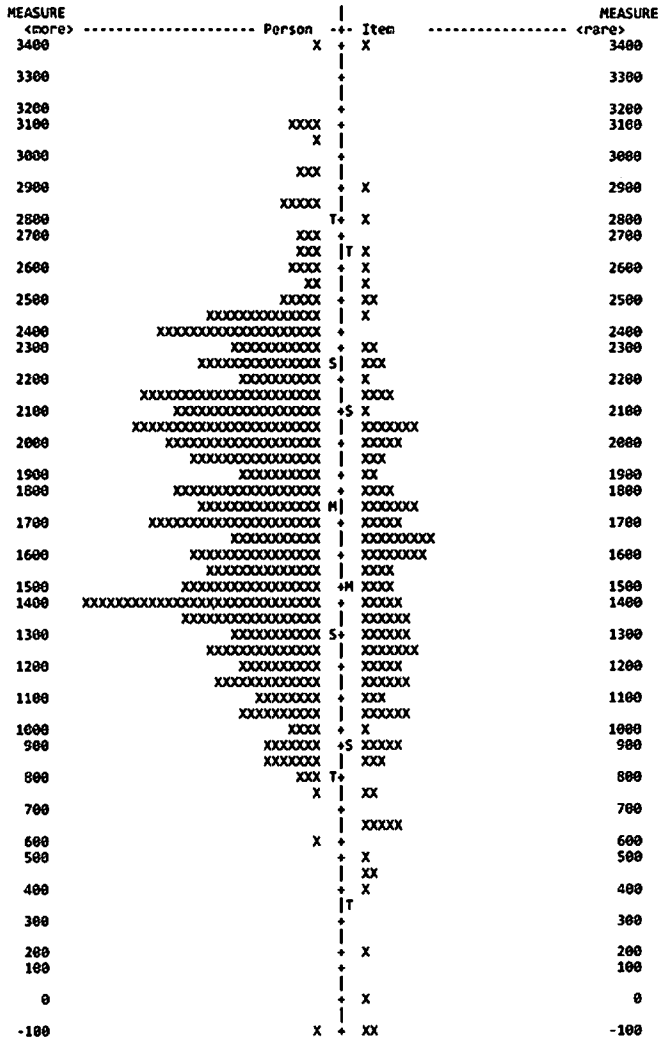

Figure 5 **Measurement rulers showing person ability and item difficulty for revised multiple-choice forms. The distribution of items and persons is somewhat mismatched, with too many easy items for this sample of persons.**

Table 5 Summary of 455 Measured Persons

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
M	27.7	48.3	1791.07	178.50	1.00	0.0	1.00	0.0
SD	9.2	5.6	494.75	41.49	0.13	0.9	0.36	0.9
Max.	47.0	50.0	3130.16	540.91	1.47	2.6	6.03	3.6
Min.	7.0	18.0	584.00	152.03	0.53	–2.9	0.25	–2.6

Real RMSE 187.97 True *SD* 457.65 Separation 2.43 Person Reliability .86
Model RMSE 183.26 True *SD* 459.56 Separation 2.51 Person Reliability .86
SE Of Person Mean = 23.22

Table 6 Summary of 146 Measured Items

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
M	87.2	152.4	1500.00	146.82	1.00	0.1	0.99	0.1
SD	62.0	117.8	588.07	90.75	0.07	1.1	0.16	1.1
Max.	255.0	323.0	3389.38	508.65	1.23	3.2	1.56	3.0
Min.	7.0	54.0	–97.92	60.92	0.85	–3.4	0.55	–2.9

Real RMSE 174.26 True *SD* 561.66 Separation 3.22 Item Reliability .91
Model RMSE 172.61 True *SD* 562.17 Separation 3.26 Item Reliability .91
SE Of Item Mean = 48.84

point-measure correlations less than .20, and the lowest correlation is .13, so these items do not seriously compromise the overall test functioning but should be excluded from operational tests.

Table 7 also lists the item fit statistics, with the mean values of infit and outfit 1.00 and 0.99 respectively, with respective standard deviations of 0.07 and 0.16. Thus item infit exceeding 1.13 and outfit exceeding 1.31 are two standard deviations greater than the mean, warranting investigation. Six items were of concern due to infit and two items due to outfit based on these criteria. Given an item pool of 150 items, this represents approximately 5% of the total, a level consistent with chance. Additionally, the substantive size of the misfit is small. Infit statistics, being information weighted, are the more crucial indicator of the quality of measurement, and the largest infit value, 1.23, represents only 23% more noisiness than the expected value of 1.00, insufficient to compromise overall measurement.

Figure 6 shows the mean vocabulary level by group and distribution of plus and minus two standard deviations. Groups 1 to 8 represent FWU students, Group 9 represents Holster and DeLint's (in press) original sample from a predominantly

Table 7 Item Statistics by Correlation Order

Item Number	Total Score	Total Count	Measure	Model SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-Measure Corr.	Pt-Measure Exp.	Exact Obs%	Match Exp%	Displace
85	45	77	1887.62	123.40	1.23	2.8	1.28	2.6	.13	.30	54.7	64.8	5.90
145	31	58	2034.67	138.47	1.14	2.0	1.18	2.1	.17	.27	51.8	60.1	4.10
89	40	77	2036.96	121.53	1.15	2.1	1.15	1.8	.19	.29	53.3	62.4	5.92
11	75	322	2319.53	71.83	1.21	2.6	1.26	2.1	.19	.37	74.4	78.9	1.24
63	24	80	2617.54	128.15	1.06	0.6	1.08	0.6	.20	.25	70.5	71.5	4.69
141	35	58	1878.84	141.31	1.11	1.3	1.10	1.0	.20	.27	51.8	63.4	4.11
94	25	76	2480.15	129.40	1.09	0.9	1.16	1.2	.20	.28	70.3	68.9	5.96
146	42	58	1576.28	155.36	1.13	0.9	1.13	0.8	.21	.29	71.4	74.0	4.11
124	25	58	2264.72	139.44	1.06	0.9	1.13	1.5	.21	.26	58.9	62.0	4.09
88	7	77	3389.38	215.18	1.03	0.2	1.56	1.2	.21	.28	92.0	91.9	6.10
117	46	55	1298.50	195.58	1.13	0.6	1.56	1.6	.21	.33	84.9	85.0	5.65
98	52	77	1664.89	130.14	1.13	1.2	1.16	1.2	.22	.31	62.7	70.7	5.88
71	49	77	1763.25	126.57	1.11	1.1	1.08	0.7	.24	.30	58.7	67.6	5.89
18	131	323	1835.79	61.97	1.16	3.2	1.20	3.0	.24	.38	62.0	68.9	1.14
123	26	58	2225.98	138.85	1.03	0.5	1.02	0.3	.24	.26	58.9	61.2	4.09
97	35	77	2184.46	121.72	1.07	1.0	1.05	0.5	.24	.29	57.3	63.0	5.93
51	57	80	1648.89	129.55	1.03	0.3	1.00	0.1	.25	.27	70.5	72.3	4.55
116	33	55	1983.13	145.04	1.04	0.6	1.04	0.4	.25	.28	58.5	62.7	5.54
130	46	58	1364.06	172.60	1.07	0.4	1.18	0.8	.25	.31	78.6	80.7	4.11
127	24	58	2303.87	140.22	1.01	0.2	1.02	0.3	.25	.26	64.3	62.8	4.09
118	12	54	2919.06	174.93	1.02	0.2	1.07	0.4	.25	.28	78.8	79.1	5.29
76	57	77	1485.32	139.22	1.09	0.7	1.11	0.6	.26	.31	70.7	76.0	5.85
96	40	77	2036.96	121.53	1.04	0.5	1.08	1.0	.26	.29	64.0	62.4	5.92
114	40	55	1664.10	159.98	1.05	0.4	1.03	0.2	.27	.29	73.6	73.8	5.61
13	136	321	1790.86	61.73	1.12	2.6	1.16	2.6	.27	.38	64.6	68.2	1.13
70	45	80	2017.83	118.11	0.99	-0.2	0.97	-0.3	.27	.26	61.5	61.4	4.63
66	55	81	1739.81	124.77	1.00	0.1	0.97	-0.2	.27	.26	69.6	69.2	4.57
45	156	311	1618.33	62.31	1.15	3.2	1.16	2.4	.27	.39	60.2	67.0	0.80
119	39	55	1714.01	156.73	1.04	0.3	1.01	0.1	.27	.29	71.7	71.9	5.60
65	29	79	2443.31	122.41	0.96	-0.4	0.99	0.0	.27	.25	66.2	65.8	4.65
144	40	58	1668.96	149.81	1.02	0.2	1.02	0.2	.27	.28	69.6	70.6	4.11
48	114	309	1948.76	65.24	1.16	2.7	1.22	2.8	.27	.40	64.2	71.6	0.85
49	219	309	1094.40	67.26	1.05	0.9	1.21	1.8	.28	.34	71.7	72.7	0.74
67	57	81	1676.38	127.60	1.00	0.1	0.94	-0.4	.28	.26	67.1	71.5	4.57
93	42	77	1977.72	122.03	1.01	0.2	1.03	0.4	.28	.30	62.7	63.1	5.91
133	46	58	1364.06	172.60	1.05	0.3	1.02	0.2	.28	.31	78.6	80.7	4.11
106	44	55	1437.51	179.43	1.04	0.3	1.05	0.3	.29	.31	81.1	81.3	5.64
7	92	323	2157.91	67.35	1.10	1.5	1.16	1.7	.29	.38	72.9	75.3	1.21
90	48	77	1794.94	125.62	1.02	0.3	1.01	0.2	.29	.30	68.0	66.8	5.89
31	213	311	1162.01	65.78	1.05	0.9	1.16	1.6	.29	.35	71.8	71.1	0.74
108	15	55	2758.89	160.94	0.97	-0.1	0.93	-0.3	.29	.27	75.5	74.5	5.35
68	45	80	2016.71	118.15	0.95	-0.7	0.95	-0.6	.29	.26	66.7	61.5	4.59
139	37	58	1797.70	143.95	0.97	-0.2	0.96	-0.3	.30	.28	66.1	65.8	4.11
92	60	77	1363.11	147.43	1.05	0.3	0.97	-0.1	.30	.32	77.3	79.3	5.83
36	204	310	1234.84	64.75	1.05	1.0	1.15	1.5	.30	.35	67.2	69.6	0.75
57	69	81	1188.80	165.82	1.02	0.2	0.90	-0.2	.30	.30	86.1	86.2	4.52
110	21	55	2477.43	146.79	0.94	-0.6	0.91	-0.8	.31	.27	66.0	65.3	5.41
96 BEST PERFORMING ITEMS EXCLUDED FOR BREVITY													
M	87.2	152.4	1500.00	146.82	1.00	0.1	0.99	0.1			75.8	76.1	
SD	62.0	117.8	588.07	90.75	0.07	1.1	0.16	1.1			11.3	10.6	

male private industrial university, while Group 10 represents a sample from a private women's university drawing from the same demographic group as FWU. Group 6 was not sampled fully due to time constraints, so the standard deviation for this group may be exaggerated by a small number of outliers. From Figure 6 we can see that the mean and distribution of FWU students is comparable to Group 10 and that the two women's universities are approximately one standard deviation above the industrial university, consistent with expectations. However, it is also apparent that most FWU groups are not statistically distinct. Although Table 5 reports person separation of 2.43, supporting the claim that the highest persons are separable from the lowest, this statistic is sample dependent. If only FWU students are sampled, the highest performing persons in Group 1 can be separated from the lowest in Group 8, but there is a large middle category of average students who are not statistically separable.

As explained by Wright (1996), separation is improved by reducing

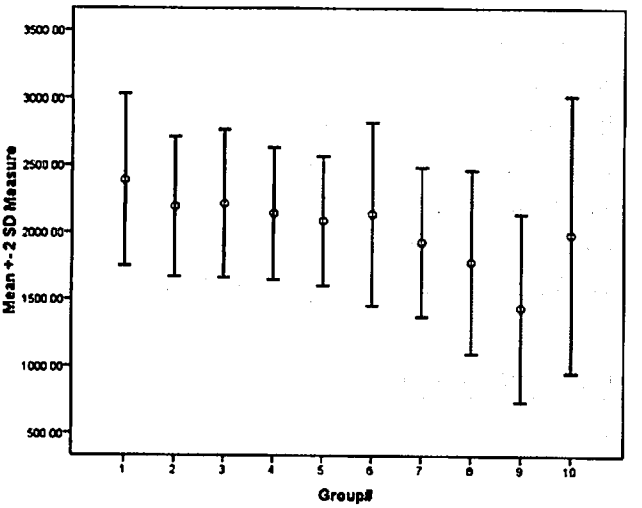

Figure 6 Vocabulary sizes by group. Groups 1 to 8 represent FWU students, Group 9 is a representative sample from a predominantly male private Japanese university, Group 10 is a sample of students from a private women's university similar in nature to FWU. The central circle represents the group mean, while the upper and lower bars represent two standard deviations from this value.

measurement error as a proportion of observed variance. This can be achieved by utilizing more items or selecting items that are closely matched to the ability of the sample of persons. Therefore, the item difficulty from version 1 and version 3 were compared and the most difficult item from each of the 50 pairs was selected for the operational test form, after exclusion of poorly performing items. Due to time constraints, piloting of this operational form was only possible with one class of 29 students, but estimations of the item performance and mapping of item difficulty against person ability are possible by combining this new data with the existing responses from these 50 items, with person ability anchored using the values from the first pilot administration. Table 8 shows that the intention of producing a more difficult form was successful, with mean item difficulty substantively higher at 1778 versus 1500 for the original.

Figure 7 maps FWU students against the item difficulty for the operational test form and we can see that, although the range of item difficulty is similar to that of person ability, mean item difficulty is still somewhat lower than mean person ability. This means that more precise measurement will occur for students of average and below average ability than the highest ability students, as can be seen in Figure 8. Measurement error is smallest when persons achieve a 50% success rate, around the 1800 word level for this test form, but measurement becomes increasingly imprecise after the 3000 word level as there are no items measuring above this level. This test provides useful measurement from the 1000 word level to the 3000 word level and will be effective in identifying learners in need of remediation but less effective at discriminating between the highest ability students at

Table 8 Summary of 50 Operational Items

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
M	58.8	94.4	1778.82	132.37	0.99	0.0	0.99	0.1
SD	19.8	3.4	613.27	41.06	0.07	0.8	0.16	1.0
Max.	90.0	99.0	2954.08	360.91	1.14	2.5	1.58	3.1
Min.	15.0	91.0	94.37	104.22	0.86	−1.5	0.67	−1.5

Real RMSE 140.05 True *SD* 597.07 Separation 4.26 Item Reliability .95
Model RMSE 138.60 True *SD* 597.41 Separation 4.31 Item Reliability .95
SE of Item Mean = 87.61

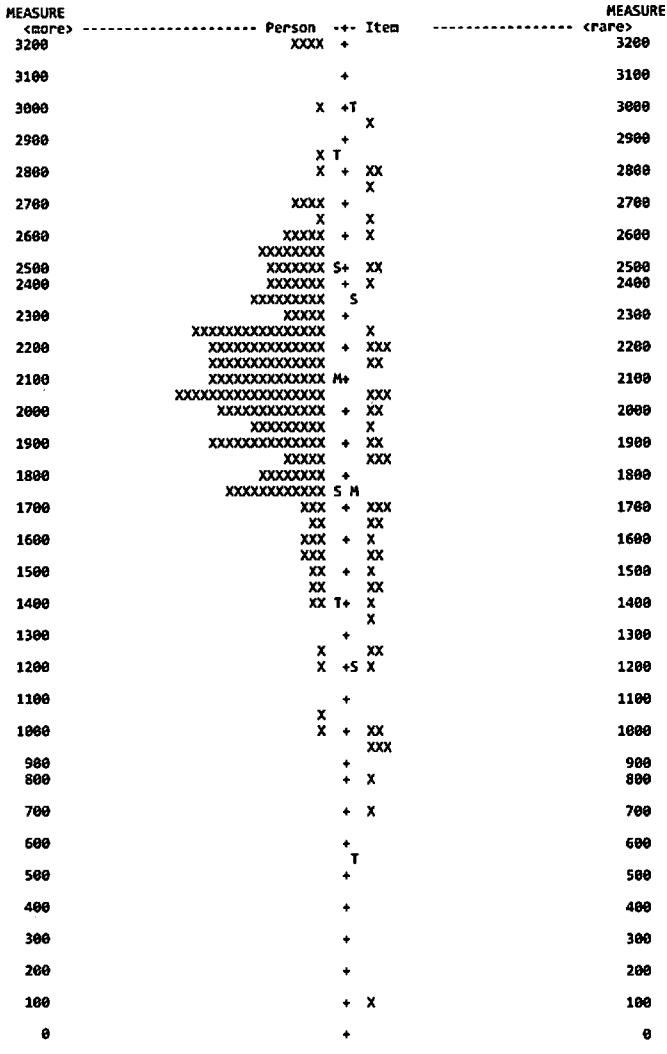

Figure 7 **Person-item map of revised test form. The distributions of item difficulty of the operational test form and target sample of persons are mapped on a common measurement scale. Compared with the distribution in Figure 5, the mean item difficulty is substantially higher, providing better targeting of item difficulty on the target population. However, the very highest proficiency students are beyond the range of measurement of this test form.**

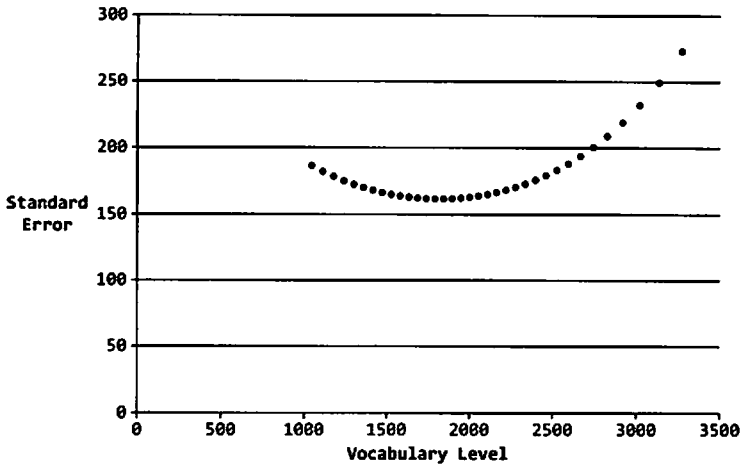

Figure 8 Measurement error versus vocabulary level. The size of measurement error of persons is shown against vocabulary level. Measurement precision is greatest near the mean item difficulty, and increases dramatically as students approach extreme scores.

FWU.

Conclusions and Future Development

Placement testing is a time-consuming process that should not be undertaken lightly. To get the most out of a placement test, it needs to be developed in the context of program and curriculum goals. It should be related to the process of needs analysis, goals and objectives, materials, other language testing in the program, and the teaching and learning going on in the institution. However, a placement test can give program administrators and teachers valuable information that they may otherwise not receive. Better information helps teachers help students and makes teaching and learning more efficient. This information also informs teachers about what is likely to be already known and boring and what is likely to be too difficult and frustrating, helping students stay motivated and gain confidence. However, in the absence of a detailed curriculum, test designers must rely on universal features of lexico-grammar that are relevant to any language

program. Items types such as cloze are a simple and practical solution in such contexts.

The use of Rasch measurement allows detailed item analysis and selection of items that provide information at the levels most relevant for the decisions being made. The operational test form developed for the FWU Academic English Program provides effective measurement for the normal range of FWU learners and can identify learners needing remediation. However, with further clarification of AEP objectives, future efforts can focus on test sections tailored more closely to the AEP program and thus achieve the full benefits that placement tests can provide for improving instruction.

Although the LVL derived cloze items address vocabulary, they do so in an integrative manner, so a synonym matching test sampling up to the 5000 word level was developed to provide better estimates of knowledge of discrete word meaning. This test was piloted at the end of semester 1 of 2011, and preliminary analysis showed acceptable levels of reliability and fit to the model, but detailed analysis will not be complete until the second semester. Given the importance of listening in the TOEFL (ETS, 2008), a listening section is essential for the AEP placement test battery. Henning, Gary, and Gary (1983) described difficulties with traditional listening comprehension tests and advocated the "listening recall" format, a listening cloze procedure. This format has the benefits of practicality and content validity, and the successful results reported by Holster (2008) make it a strong candidate for inclusion. Sample items have been written and it is expected that pilot administration of the full test form will be completed in semester 2 of 2011. Reading and writing are key components in the AEP, so testing ability with paragraph length written texts is also necessary. The cloze-elide procedure, where candidates must identify extra words inserted into a text as a test of proof-reading has been found to be a practical measure of general language processing, so this will be investigated in semester 2 of 2011 as a candidate for a test of written language knowledge.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

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

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level vocabulary tests. *Language Testing, 16*(2), 131-162. doi: 10.1177/026553229901600202

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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum Associates.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 65-83.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston: Heinle & Heinle.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Brown, J. D. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies, 21*, 79-125.

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

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

Chauncey, H., & Frederiksen, N. (1951). The functions of measurement in educational placement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 85-116). Washington D.C.: American Council on Education.

Clark, M. (2004). By the numbers: The rational for Rasch analysis in placement testing. *Second Language Studies, 22*, 61-90.

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

ETS-IIBC. (2002). *TOEIC Bridge guide and questions*. Tokyo: IIBC.

ETS. (2008). The TOEFL® Test - Test of English as a Foreign Language™. Retrieved 28 March, 2008, from http://tinyurl.com/zocgc

Field, A. P. (2009). *Discovering statistics with SPSS* (3rd ed.). London: Sage.

Henning, G., Gary, N., & Gary, J. O. (1983). Listening recall: A listening comprehension test for

low proficiency learners. *System, 11*(3), 287-293.

Holster, T. A. (2008). A reliability analysis of summative cloze test formats. Retrieved 25 May, 2008, from http://www.sun.ac.jp/sangaku/study/pdf-m/kiyou003.pdf

Holster, T. A., & DeLint, D. F. (in press). Output tasks and vocabulary gains. *The Language Teacher*.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44. doi: 10.1093/applin/19.1.24

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language, 17*(1), 23-59.

Inoue, N. (2006). What's going on inside the pine tower of babel: Foreign language curriculum reform in a Japanese university. *Languages and Cultures Series, 16*, 87-115.

Jonz, J. (1976). Improving on the basic egg: The M-C cloze. *Language Learning, 26*(2), 255-265. doi: 10.1111/j.1467-1770.1976.tb00276.x

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal, 73*, 440-464.

Lake, J., & Holster, T. A. (2011). *Vocabulary test format effects*. Paper presented at the 10th JALT Pan-Sig Conference, Shinshu University, Nagano.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Linacre, J. M. (2010). *A user's guide to Winsteps* (Version 3.70.02). Retrieved 19 September, 2010, from http://www.winsteps.com/winman/index.htm?copyright.htm

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Pearson Education.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied Linguistics in Society*. London: CILT.

Mukherjee, J. (2001). Principles of pattern selection: A corpus-based study. *Journal of English Linguistics, 29*(4), 295-314.

Pienemann, M. (1984). Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition, 6*(02), 186-214. doi: 10.1017/S0272263100005015

Pienemann, M., Brindley, G., & Johnston, M. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition, 10*(2), 217-243.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88.

Sinclair, J. M. (Ed.). (2004). *How to use corpora in language teaching*. Amsterdam: John Benjamins.

Stahl, S. A. (1999). *Vocabulary development*. Cambridge, MA: Brookline Books.

Szabo, G. (2008). *Applying item response theory in language test item bank building* (Vol. 10). Frankfurt: Peter Lang.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 414-438.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3), 321-344. doi: 10.1177/026553229401100305

Westrick, P. (2005). Score reliability and placement testing. *JALT Journal, 27*(1), 71-92.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics, 21*(4), 463-489. doi: 10.1093/applin/21.4.463

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20*, 1-28.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions, 9*(4), 472.