Research Paper

Validating Classroom Assessments Measuring Learner Knowledge of Academic Vocabulary

John Rylander*, Catherine LeBlanc, David Lees, Sara Schipper, and Daniel Milne

Abstract

This research investigates the reliability of vocabulary assessments used in English language skill classes at Kyoto University, Japan. Specifically it examines assessments designed to measure learner knowledge of a set of 477 vocabulary items from the *Kyoto University Academic Vocabulary Database* （京大学術語彙データベース：基本英単語1110） in the first-year curriculum of a combined skills writing/listening course. Learners provided responses to two item types—multiple-choice and fill-in-the-blank—in an online quiz format constructed using Google Forms and delivered using QR codes. Results show that the vocabulary instruments revealed good model fit based on a Rasch analysis of each quiz version. Follow-up statistical analyses were performed in the form of *t* tests. These analyses, which used Rasch logits of item difficulty, revealed that items measuring receptive knowledge showed lower item difficulty than those measuring productive ability. A person-level analysis of gains over time showed somewhat mixed results, with just two of the five sample populations revealing gains between spring and fall semesters. This analysis was performed as part of a larger needs analysis project with results providing insights into how best to structure additional in-house created materials.

[Keywords] Needs analysis, vocabulary acquisition, English for General Academic Purposes (EGAP), Rasch analysis

Research into second language (L2) vocabulary acquisition has investigated the relative effectiveness of various learning methods (Mondria & Mondria-De Vries, 1994; Nation, 1997; Nation, 2001), provided descriptions regarding the nature of vocabulary itself (Leki & Carson, 1994; Milton, 2009; Wilkins, 1972), and viewed the circumstances under which vocabulary is acquired and used (Krashen, 1989; Melka, 1997; Mochizuki, Aizawa & Tono, 2003; Nation, 1990, 2001, 2010; Schmitt, 2010; Schmitt & McCarthy, 1997). Furthermore, the structure, creation, and utilisation of academic word lists have also featured heavily in L2 vocabulary research (Chung & Nation, 2003; Coxhead,

* Program-Specific Senior Lecturers, International Academic Research and Resource Center for Language Education

2000; Coxhead & Nation, 2001; Martin, 1976). The literature review to follow covers: (a) aspects of vocabulary—including knowledge, frequency, and links with language proficiency, (b) features of English for Academic Purposes (EAP) vocabulary, and (c) the use of L2 academic vocabulary in Japanese post-secondary institutions.

Vocabulary Knowledge

Although there currently is no overarching, united theory of how lexical knowledge is acquired (Schmitt, 2010), there are commonly agreed upon concepts concerning vocabulary knowledge and acquisition in general. First, concerning L2 vocabulary size, research suggests that while 6,000–7,000 word families are sufficient for informal daily conversation, native speaker university graduates know closer to 15,000 word families (Schmitt, 2010). Achieving such a broad vocabulary requires a considerable time commitment, both in and out of classes (Milton, 2009), as well as a large amount of conscious and deliberate interaction with the L2 (Krashen, 1989; Schmitt, 2010). In the L1, academic vocabulary is far more complex than more commonly used lexis, due to the technicality of their use and meaning and correspondingly low frequency. Academic vocabulary for L2 learners, especially for university undergraduates, can prove challenging, as most must expand and focus their command of academic vocabulary beyond those taught as part of the secondary education curriculum (Leki & Carson, 1994). Additionally, without regular revisiting, vocabulary knowledge attrition will likely occur (Milton, 2009). That said, previously learned vocabulary is more quickly reacquired than those learnt anew (Schmitt, 2010), and incidental vocabulary exposure can be facilitative in this endeavour (Krashen, 1989).

Regarding depth of vocabulary knowledge, it is generally held that in order to correctly 'know' a word a learner must know the spoken form, the written form, its meaning, associated grammatical patterns and words, as well as its frequency and register (Nation, 2001; Schmitt, 2010). Schmitt (2010) reports that vocabulary learning is incremental in nature mainly because these individual types of word-knowledge are acquired separately and over time. Additionally, Milton (2009) notes that these word-knowledge types also have cognitively discrete 'receptive' and 'productive' qualities; such knowledge naturally acquires and atrophies at different rates. Given this incremental nature, its varied methods of acquisition, and the tendency for EFL learners' vocabulary knowledge to be more receptive than productive (Melka, 1997), word knowledge is likely to differ depending on the actor and perspective. Research conducted into Japanese university students suggests that average vocabulary sizes range from 2,000 words to 4,000 words (Mochizuki, Aizawa & Tono, 2003). This raises the issue of how best to increase the size of vocabulary knowledge across cohorts of learners.

Word Frequency

Word frequency, which is generally defined as the number of word occurrences in a given corpus, is also an important aspect of understanding vocabulary acquisition. Nation (1997) notes that when considered alongside the vocabulary size of L2 learners, a frequency-based list of words "provides a rational basis for making sure that learners get the best return for their vocabulary learning effort" (p. 17). For skills associated with academic fluency (e.g., the presenting, discussing, writing,

and reading of scholarly works), the technical, specialised vocabulary required for effective communication within specific disciplines is usually of a lower frequency. A second point for consideration is the specialised vocabulary required for certain fields (Martin, 1976). When studying literature, one might encounter words like "idiosyncratic" or "enlightenment", and in more scientific disciplines uncommon words such as "quadratic" and "specimen" may need to be understood. Finally, given the relative obscurity of lower frequency words, it is often difficult to know which words to target, at least for EAP, and the best strategies for how to acquire them. Under such circumstances, explicit and deliberate vocabulary practice focused on a select list of words may prove useful for L2 learners (Nation, 1997).

Vocabulary Knowledge and Language Proficiency

Previous research suggests a strong correlation between vocabulary knowledge and general language proficiency. Linked with Wilkin's noting that without vocabulary nothing can be conveyed in an L2 (1972) and the fact that "words are the basic building blocks of language" (Read, 2000, p. 1), it appears that vocabulary size and grades on general tests to measure language-ability are positively correlated (Milton, 2009; Schmitt, 2010). While it should be noted that "language level is not just knowing language knowledge, but using it communicatively", when it comes to language assessments (which tend to rely on written, receptive knowledge) a broader, deeper vocabulary tends to lead to higher scores (Milton, 2009, p. 171). Additionally, phonological vocabulary knowledge also correlates relatively well with listening test scores (Milton, 2009).

Academic Vocabulary/EAP: ESAP-EGAP

Academic vocabulary is tentatively defined as "low frequency, context independent words occurring across disciplines" (Martin, 1976). Several researchers suggest these words could, at their most discipline-specific, be called "technical vocabulary" (Chung & Nation, 2003; Coxhead & Nation, 2001) and refer to English for Specific Academic Purposes (ESAP) words. These words are considered to be discipline specific terms and are somewhat difficult to encapsulate. A slightly broader definition, covering much of the overlapping ESAP words from multiple disciplines, might consist of words termed by researchers as "academic vocabulary" (Coxhead, 2000; Martin, 1976) or "semi-technical vocabulary" (Farrell, 1990) and can be roughly labelled English for General Academic Purposes (EGAP).

Such difficulty in specifying which words can be considered generically applicable in academic contexts has resulted in wordlists such as the Academic Word List (AWL) (Coxhead, 2000). Coxhead (2000) suggests that EAP vocabulary consists of non-general words that do not appear in the General Service List (GSL; West, 1953). These lists, and more recently the New Academic Word List (NAWL) based on the New General Service List (NGSL; Browne, Culligan, & Phillips, 2013), continue to serve as the basis for an extensive number and variety of course types, as well as for department-created lists. Based on surveys of academic papers and documents, 78.2% of the words were included in the GSL and 8.5% were from the AWL (Nation, 2001); this leaves roughly 13.3% of the total words in the surveyed texts unaccounted for by these two lists combined. This suggests that, given mastery of the vocabulary in these two word-lists, L2 learners should be able to deal with academic texts of this lex-

ical make-up. Tajino, Stewart and Dalsky (2010) suggest that finding ways to add effective academic vocabulary training is of substantial importance for advancing learner awareness and productive ability of academic vocabulary.

Assessment of Productive and Receptive Vocabulary Knowledge

Research into the assessment of vocabulary knowledge makes a distinction between learner receptive vocabulary knowledge and productive knowledge (e.g., Fitzpatrick & Clenton, 2017). Multiple-choice or matching questions are often used to measure the former and have been used as part of the Vocabulary Levels Test (VLT), originally conceived of by Nation (1983) and subsequently revised numerous times (e.g., Beglar & Hunt, 1999; Nation, 1990; Schmitt, 2010). In general, multiple-choice questions are relatively easy to create, administer, and score. Moreover, bias in teacher scoring can be avoided, as only one answer is considered correct (Brown & Hudson, 1998). In regards to measuring productive vocabulary knowledge, short answer or fill-in-the-blank questions are common, which can be found in the vocabulary-size test of controlled productive ability (VTCPA), developed by Laufer and Nation (1999). Fill-in-the-blank questions are described as measuring 'controlled productive ability', in that they only require the learner to have the ability to produce a word in a constrained context (Laufer & Nation, 1999 p. 37). Despite the limited context, Laufer and Nation (1999) found that providing the first two or more letters of a fill-in-the-blank answer to help distinguish it from other potential items being tested could potentially provide a valid method for distinguishing students of differing proficiency levels. To measure both receptive and productive vocabulary knowledge, employing multiple tasks will generally lead to improved assessment validity (Brown & Hudson, 1998).

Situation of Academic Vocabulary in Kyoto University

Based on the concerns discussed above, a collection of academic vocabulary was created for use by students enrolled in general studies courses in their first two years at Kyoto University. This database, known as the *Kyoto University Academic Vocabulary Database*, was compiled and constructed based on EAP principles for use in a newly designed first-year English (EGAP) curriculum. Tajino, Dalsky, and Sasao (2009) state that the EGAP vocabulary list would be positioned where ESAP vocabulary requirements from the Agriculture, Literature, Philosophy, Economics, Law, Education, Medicine, Pharmacy, Science, and Engineering departments overlap. Having established this, they note that the question becomes: "What kind of vocabulary is appropriate for what kind of group of students at what stage?" (p. 9). While there may be several concerns about EGAP not being as specifically cost-effective as more focused ESAP, the list creators considered that an amalgamated EGAP word-list was reasonable, as the vocabulary in question is designed for inclusion in a year-long course.

As an initial step in creating the Kyoto University database, 1,651 research articles from the Agriculture, Literature, Philosophy, Economics, Law, Education, Medicine, Pharmacy, Science, and Engineering disciplines were selected at random from a list of recommended research journals provided by academics across faculties. The articles were then converted to data format, lemmatized,

and amalgamated into a corpus (the Kyoto University Academic Research Article Corpus). By cross-referencing data from the GSL to exclude the more frequent, general words, Tajino, Dalsky, and Sasao (2009) report that though ESAP vocabulary varies between disciplines, it can be categorized into four subgroups: (a) Medicine, Agriculture, and Pharmacy; (b) Economics, Law, and Education; (c) Science and Engineering; and (d) Literature. These subgroups can be further categorised into three groupings: (a) Science, (b) Medicine and Biology, and (c) Humanities and Arts. When organized in this manner and then introduced into a curriculum that includes instructional content across academic fields, EGAP (General), EGAP-A (Literature/Arts), and EGAP-S (Science) word lists might prove beneficial as a means of supporting deliberate vocabulary study. The authors note that after curriculum developers fully understand the facts of the learning situation, word lists can then be "developed by pedagogically determining the selection criteria" (Tajino, Dalsky, & Sasao, 2009, p. 17).

Research on the Kyoto University Academic Research Article Corpus has led to the development of the *Kyoto University Academic Vocabulary Database*, a list of 1110 words separated into EGAP, EGAP-A and EGAP-S terminology. The words for the vocabulary tests created for the current research project were drawn from the first 477 found in the general EGAP section of this list.

Overview and Purpose of the Present Study

In the field of second language acquisition (SLA), a needs analysis approach provides curriculum designers a method for investigating the efficacy of materials under review. Most commonly, this approach has been used in English for Specific Purposes (ESP) courses. One benefit a needs analysis serves is to identify the role specific goals and objectives play in a newly formed or established curriculum and set of instructional practices. Researchers using a needs analysis as a framework for curricular review employ a deliberate and systematic data collection procedure to assess a set of clearly defined needs. From this, researchers can then consider the relative effectiveness of materials and assessment procedures; findings from these investigations can then be applied during creation of new components of the curriculum creation or enhancement of portions in the curriculum requiring emendation or alteration. In conducting a needs analysis, faculty researchers generally frame a particular need in terms of the discrepancies apparent in an existing program or by concentrating on ways of enhancing a particular knowledge or skill deemed essential for a particular set of learners. In this way, a needs analysis approach technically constitutes an information-gathering process performed as part of routine, in-house diagnostic review of how stated goals are achieved.

Therefore, the main purpose of this study is to evaluate the reliability and validity of assessment tools designed to measure a vocabulary component built into a first-year academic writing and listening course. The explicit study of English vocabulary for general academic purposes from the *Kyoto University Academic Vocabulary Database*（京大学術語彙データベース：基本英単語1110） as part of the course curriculum led to the creation of a set of quizzes to assess learning and, in turn, to evaluate of the extent to which curricular goals relating to learner vocabulary development were achieved as outlined in the course syllabus. These quizzes, produced in-house by a team of curriculum designers and classroom lecturers, are the only aspect of the curriculum at present customized to fit student learning. Thus, a systematic analysis of data collected over one academic year was

performed to determine: (a) the reliability of the instruments (i.e., the five sets of quizzes), (b) the level of difficulty between item types, and (c) learning gains over spring and fall semester. Following a principled approach in the development, evaluation, and maintenance of the course curriculum, results from this research will lead to the revision and improvement of quizzes and will inform future curricular decisions regarding materials development across other areas of the course.

Methods

Participants

Participants represented first-year Kyoto University students enrolled in compulsory English Writing and Listening (EWL) classes for the spring and fall semesters of the 2016 school year. This study analysed a subsample of 39 of the total of 148 EWL classes of that year. Administrative planning for these courses results in sections of approximately 40 students from the same faculty being divided into two groups, each composed of about 20 students. One instructor was assigned to each group, and from spring to fall these groups switched instructors. English proficiency levels played no part in student streaming or scheduling. The faculties, along with the number of classes from each faculty covered in the study, were: Engineering (15), Letters (4), Law (4), Science (4), Economics (3), Medicine (3, including 1 from Health Science), Integrated Human Studies (2), Pharmacy (2), Agriculture (1), and Education (1).

The EWL classes focus on developing academic writing and listening skills and are part of compulsory courses worth two credits each semester spanning one academic year, with both semesters composed of 16 lessons of 90 minutes scheduled as 14 periods for instruction, one for examination, and one for individual student feedback. Vocabulary assessments made up 10% of the overall course grade for both the spring and fall EWL courses, with the bulk of course grades allocated to writing assignments (60% in the spring and 50% in the fall semester), listening assessments (which combined made up 30% in the spring and 20% in the fall), and a year-end performance on a standardized English language proficiency test (20% of the fall grade).

Materials

The vocabulary assessments took the shape of nine quizzes composed of five multiple-choice items and five fill-in-the blank items. In sum, the quizzes represented 90 items, with 45 items representing each item type. When initially constructing these quizzes, 477 words in the general part of *Kyoto University Academic Vocabulary Database* were divided into eight sections consisting of 50 vocabulary items each and one of 77 items. For test security, five different versions of each quiz were created and then assigned a letter from A–E. Random vocabulary items were selected from specific sections of the database, with individual items representing the various versions. This reduced item overlap across quiz versions and item types. Each version was then given to learners on a different day. For instance, students taking Monday lessons would take all nine of the A quizzes across the duration of the semester; Tuesday students would take all B quizzes. In the fall, the same 477 words were tested once agein; however, the quizzes were rotated so that Tuesday students would take the A

version of the quizzes and Wednesday students would take the B version quizzes.

As mentioned, each quiz consisted of 10 questions (five multiple-choice and five fill-in-the-blank) and students were given roughly seven minutes to complete each quiz. For the multiple-choice questions, a definition was given, and students were instructed to choose one correct answer from four possible options. The randomized distractor options were of the same part of speech as the target vocabulary (see Figure 1). The second part of the quiz contained five fill-in-the-blank questions. For these questions, students were given a sentence with the target vocabulary missing. No word bank was provided, but the first letter of the target word was shown (see Figure 2).

Quizzes were written using Google Forms, which allowed for ease in the creation and editing of questions. It also enabled the teacher-researchers to share information and divide the workload. A QR code was generated for each quiz, allowing each to be administered in class electronically by projecting the code onto a screen or via an instructional monitor (see Figure 3 below). Students then scanned the code with their devices (normally smartphones or tablets) using a freely downloadable QR code reader. (Over the course of a year of data collection, only five students requested to use a paper-based version of the quizzes.)

Google Forms allowed for all learner responses to be archived (including required information regarding first and last name, student ID, and course instructor) in a Google Sheet, along with a timestamp that indicated when the learner pressed the “submit” button, down to the second. Each element of the form, including the personal information, was stored in a separate cell of the sheet,

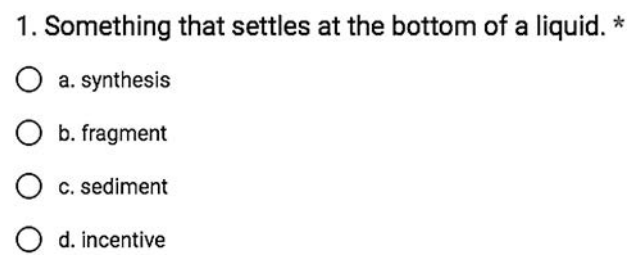

Figure 1. Multiple-choice example. This figure shows an example multiple-choice item.

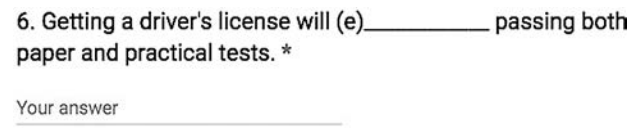

Figure 2. Fill-in-the-blank example. This figure shows an example fill-in-the-blank item.

Figure 3. Vocabulary quiz QR code (an example quiz can be accessed through this code).

which could then be downloaded for further analysis. As long as a quiz remained active, the Google sheet would continue to collect response data following user submissions. This way, record-keeping for all quiz responses over the course of a year was performed without hassle, and when the collection period concluded, each sheet was downloaded and secured in an offline form.

Rasch Model

The Rasch model (Rasch, 1960) provides a mathematical method for transforming ordinal data in the form of raw test scores into various distinct models that construct units of measurement on a true interval scale. For many analyses using the Rasch model, these units of measurement apply to two variables: persons and items. Software such as Winsteps provides a means of reviewing numeric and graphic coefficients for person and item parameters and their associated fits statistics, which allows for a closer inspection of how the compiled data fit model expectations. The dichotomous model used in this analysis is one such model that provides coefficients representative of these two variables on a linked scale. Rasch considers the variables of person and item as measurable latent traits factored into the analysis simultaneously.

The Rasch formula is expressed as $P_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$, where P_{ni} represents the probability of a person (n) with particular ability (B*n*) responding successfully to an item ($_i$) of a particular difficulty (D_i). Here, the *exp* signifies the natural constant *e*, which is set to the value of 2.71828. For persons, the interval scale represents the range of abilities displayed by the sample of individuals through an aggregate tally of responses on a set of items designed to represent a distinct construct. This corresponds to the basic assumption that when assessing individuals on any type of test the resulting scores will reveal a spread between the better performing, less challenged persons and those more challenged by the assessment. The greater an assessment can spread persons, the more beneficial for the analysis, as no two persons can fundamentally possess an identical amount of knowledge (due to such factors as intelligence, experience, practice, motivation, age). For items, the interval scale represents the range of difficulties assumed measurable across the latent trait. Here, that trait is vocabulary knowledge, which is represented in two forms: (a) receptive knowledge, as exemplified via multiple-choice items, and (b) productive knowledge, as exemplified via fill-in-the-blank items. The assumption here is that just as various individuals should display differing amounts of ability vis-à-vis any particular construct, items representing that construct will exhibit varying levels of difficulty. In the case of measuring academic vocabulary, this may occur due to the task type, such as with the item format, or may occur due to the nature of the lexical item itself.

As the data used for the vocabulary instruments reported below derive from participant response strings on either multiple-choice items, with one correct answer and three distractors, or fill-in-the-blank items, which are either correct or incorrect, we analysed the data using the Rasch dichotomous model (Rasch, 1960). The dichotomous model calculates the probable success any person having a set degree of ability on the latent trait—represented as B_n in the model equation—will have when encountering a particular item at a set level of difficulty—represented as D_i. This probability fluctuates based on the size of any person's ability estimate relative to any item's difficulty estimate. For instance, when these two estimates are equal and the distance between the two is zero (i.e., when

a person of a particular estimate encounters an item of matching difficulty), the probable success that person will answer the item correctly is fifty-fifty, or 0.5. As the distance between the person's ability and the item difficulty increases, the probability of success will do likewise, either increasing in the case of greater ability and less difficulty or decreasing in the case of less ability and greater difficulty.

The interval measure used to link the ability and difficulty estimates along a shared scale is the logit, a contracted word form of *log odds unit*. This scale is represented both numerically as well as graphically (see the Wright map shown in Figure 4) and provides a means of inspecting the assumption that within any sample a range of measurable abilities exists in relation to the construct of interest (as represented by the aggregate sum of instrument items). The linked scaling also provides a means of inspecting across persons to items, due to their interconnected relationship in the model, and determining relative instrument difficulty compared to relative person abilities (i.e., how difficult a test is for a sample population). The combined data is presented hierarchically, with definitive steps pegging relative difficulties and abilities. For instance, if a person sits a 100-item assessment and scores an 80 (i.e., 80% correct), the equation including the natural log appears as such: $ln(80/(100 - 80)) = 1.39$. If a person only scores 75, the logit ability would amount to 1.09. In contrast, consider that a single item on the same 100-item assessment is answered correctly by 60% of the sampled test-takers. The same logit conversion process results as: $ln(60/100 - 60)) = 0.4$. If the item were answered by only 20% of the test-takers, the logit of item difficulty would represent a higher estimate: 1.39. This calculation on person performance and item response forms the hierarchy of persons and items.

Preliminary Analyses

The preliminary analyses involved an investigation of participant response behaviour and item fit for the five data sets represented by Vocabulary quizzes A, B, C, D, and E using Rasch analysis via the Winsteps reports. The first analysis involved a visual inspection of the items using the Wright map, with a discussion of how the items shape the hierarchical data structure. The second analysis includes a report of the Rasch reliability and separation coefficients for persons and items, followed by a brief discussion of how the vocabulary assessment reflects the range of abilities of test-takers. The third section contains a report of the statistics showing how the items and persons fit Rasch model expectations. The fourth and final section reports on the analysis of instrument unidimensionality across the five instruments and covers issues relating to the principle components analysis (PCA) of the Rasch item residuals and the corresponding contrasts.

Wright Map

The initial visual inspection of the data included a review of the Wright item map produced using Winsteps. Wright maps may either be set to highlight the item hierarchy or that for the persons, with the focus of interest displayed to the right of the vertical line. In addition, either the person measures or item measures may be centred; here it is the latter, with an arbitrary transformation centring at 50. The example Wright map (Figure 4) shows the item hierarchy for Vocabulary quiz A

(VQA). The top of the figure provides a label to the right showing *More able persons* and to the left *More difficult items*. Conversely, at the bottom of the figure to the left is the label *Less able persons* and to the right *Easier items*. Running vertically down the figure is a dashed line denoting the separation between these split sides of the analysed data. The right half of the map is the focus of interest and reflects the items, which is where the scale is centred, while the left half reflects the cluster of test-takers. Along the centreline, in addition to the dashes (i.e., pipe symbol) falls another symbol (+), which represents the 10-point ascent/descent of the scale. Occasional breaks in the dashed line appear, which simply show where either the string of persons or items (as is the case here) from one line spill onto the line below. Where this occurs, this should be read that those persons or items from the broken dash line are considered indistinguishable from those in the line immediately above it. To the immediate right of this line are the letters *m*, *s*, and *t*. The letter *m* signifies the mean of each of the two data sets. The letter *s* signifies one standard deviation away from the mean. The letter *t* signifies two standard deviations away from the mean. These show the degree of data dispersion along the common scale and allow for a quick reference of where the mean person abilities lie in comparison to the mean item difficulties.

There are several steps to visually inspecting the data structure. The first is to consider the relative placement of the two letters *m* representing the mean abilities of the persons (on the left) and that of the item difficulties (on the right). The mean for the item difficulties is set to 50, and the person ability estimates are free to float, with it resting at approximately 58, just below the first standard deviation above the item mean. This shows that on average for the collection of items designed to measure academic vocabulary knowledge, the participants proved themselves generally capable. Second, by inspecting the letters *s* and *t* as to the data dispersion, it is clear that the range of abilities displayed by this sample appears relatively homogenous by comparison to the item difficulty estimates. The cluster of person ability estimates registers a low of approximately 47 and a high of 70, while the item difficulty extends from the zero point up to approximately 74. Third, while only item 18 (conflicts) appears beyond the ability of all persons, a total of 31 items (34%) falls below the least able person. Of these 31 items, 30 of them (97%) represent a multiple-choice item type (shown through the second digit consisting of a number between 1 to 5). The other point to consider regarding this feature is that 59 items (66%) rest horizontally across from the range of person estimates. Recall that the person ability estimates are created from the item difficulty estimates, and vice-versa. The reliability estimates produced from these pegged calculations rely on persons and items measuring at the same position. With all persons resting abreast of a cluster of items, this indicates high reliability for person measures. However, with a third of the items measuring below the lowest ability estimates for the sample, this may suggest attenuated item reliability. Fourth, the item coverage contains minimal gaps at the top and bottom and no gaps in the middle. This is significant as it indicates that the range of items representing the construct of vocabulary knowledge provides an adequate accounting of the range of person abilities measured against items. However, the density of item coverage (i.e., the number of items sharing the same horizontal location) shows that many items may have been ineffective in separating person ability, as they measure ability levels in the same manner as other items (i.e., too much item clustering). Lastly, the relative density of dispersed

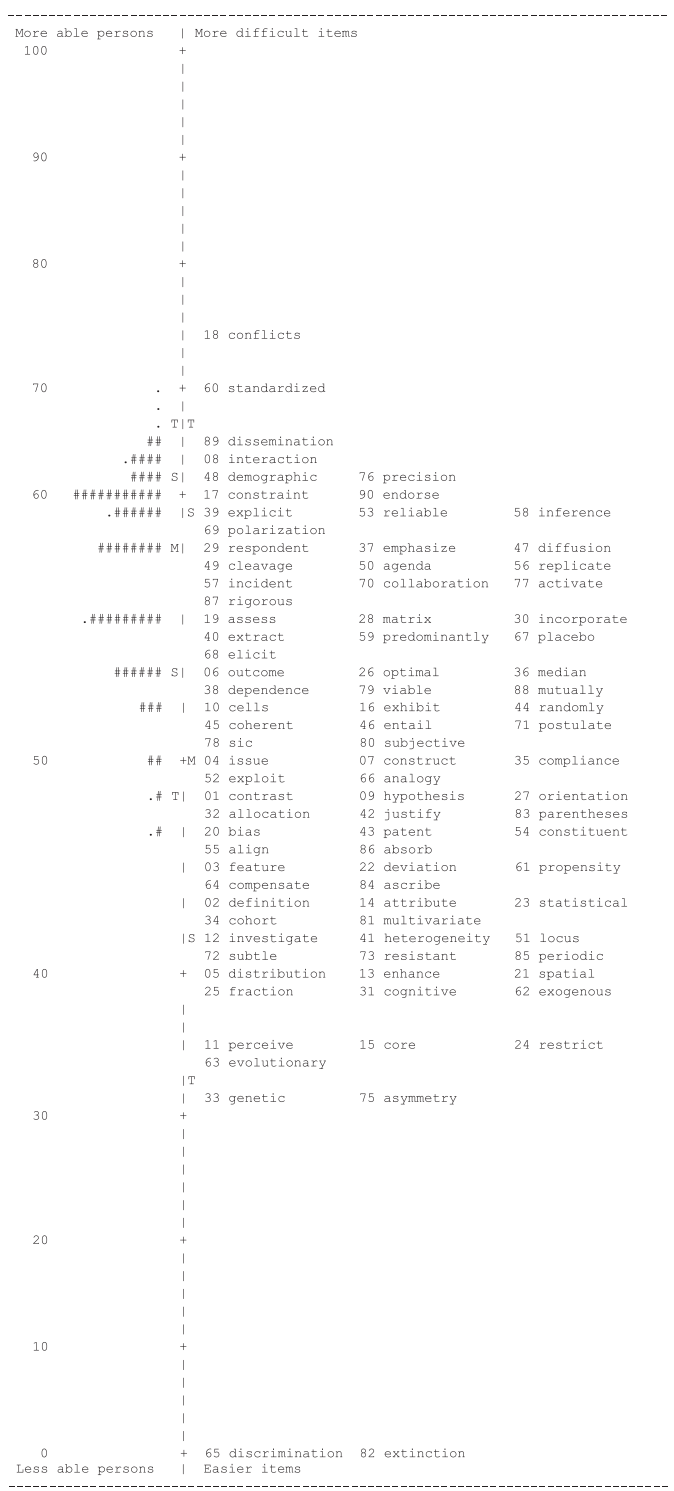

Figure 4. Vocabulary quiz A Wright item map (*N* = 188; k = 90). M, S, and T represent mean, one standard deviation from the mean, and two standard deviations from the mean, respectively. The symbol "#" signifies a cluster of two persons, and the symbol "." signifies single persons.

persons appears more compact than that of items, which is a sign of a homogenous sample in regards to item responses. A pattern of this nature suggests that the items provide more information about the persons than the persons do about the items.

With the use of the Wright map, the graphic results from the Rasch analysis provide a representation of both a norm-referenced and criterion-referenced nature. In norm-referenced form, the data reveal the relative distances between individuals along the vertical scale set to an interval measure. In relation to item difficulties, what this means, for example, is that the difference of difficulty between item number 33 (genetic) and item number 72 (subtle) is the same as between item 72 and item 78 (sic), just as the difference between 10 degrees Celsius and 20 degrees is considered the same as the difference between 20 and 30 degrees. This applies to the person difficulty data, where a person resting at 50 (which would make that person one of the least able in this group) differs in precisely the same way from a person resting at 60 than that same person at 60 and one of the few persons measured at the higher end of the ability scale at 70. Raw score data provide only ordinal representations of the response strings, and in that condition, fail to meet basic assumptions required of further statistical analyses (i.e., they represent ordinal data not interval or scale data). Likewise, the data offered in the Wright map provide a criterion-reference view of the content and persons, when looking across the horizontal between the linked data. That is, even the persons with the lowest ability estimates display some command of vocabulary knowledge over 31 items—all of which reveal difficulty estimates at the bottom portion of the map. In fact, only 12 vocabulary items appear beyond the ability of one half of the sampled population. Viewing the Wright map in this manner provides insights regarding each individual's relative acquisition of a subject matter, with items below a person's measure indicative of what that person has gained mastery of and items above that person's measure indicative of knowledge yet to be learned completely. Items immediately abreast of a person's measure best represent content in a stage of development.

The data presented in Figure 4 have been transformed from the logit scale to a CHIPs scale, which sets the mean of either items or persons to 50 rather than at 0, with a scale of 4.55 and a mid-range of 50. By transforming the data in this way, item relationships remain tethered to one another as they were in a logit scale but with a larger spread along a non-zero mean that extends from 0 to 100. The benefit of this transformation process is to remove negative values from follow-up statistical analyses and to simplify reporting across readerships.

Reliability and Separation Coefficients

The second step in investigating the data derived from the vocabulary instruments involves moving beyond the graphic representation and onto measures associated with how the data fit the model. As stated, the Rasch family of models, with the dichotomous model applied in this research, provides a method of producing a metric level of measurement at the interval level on a latent variable. To ensure the validity of this form of measurement requires satisfying the properties of sufficiency, separability, objectivity, and latent additivity (Rost, 2011). Sufficiency relates to how well the items and persons fit model expectations. Separability relates to the estimation of item parameters. Objectivity relates to the independence of the sample of items used in designing the instrument and

the participant sample sitting that instrument, which relates to the generalizability of the measures. Latent additivity relates to interconnectedness of person measures with those for items. For these to apply, the question is not whether the model fits the data, as with traditional statistical estimation, both standard and advanced (e.g., *chi-square, comparative fit index, incremental fit index, expected cross-validation index, goodness of fit index*). Rather, with Rasch analysis, the question is how well the data conform to model expectations. Confirming these expectations requires an investigation of the reported data-to-model coefficients.

In addition to the graphic representation provided in Wright map form, Winsteps reports reliability and separation estimates. As the names of these coefficients suggest, the reliability coefficient reports the degree of true variance from the total observed variance found in the data, and the separation coefficient reports the number of statistically distinct levels along the linked hierarchy. Put another way, reliability shows how confident test designers can be of the data hierarchy, so that if another similar sample of participants were to sit an instrument composed of similar items the likelihood of the same results would occur. Fisher (2007) states that Rasch reliability coefficients are similar to those reported for Cronbach's alpha and range from a low of .0 to a high of 1, with coefficients reporting a reliability at or greater than .94 as excellent, between .81 to .93 as good, from .67 to .80 as fair, and those at or below .66 as poor. In terms of separation criteria, >5 is considered excellent, from 4 to 5 is very good, from 2 to 4 is acceptable, and below 2 is poor. Winsteps reports these estimates for both persons and items.

Table 1 provides the reliability and separation estimates for all five vocabulary quizzes. The reported sample sizes for the five quizzes represent final tallies after a preliminary investigation of the data containing all response strings for all participants sitting each quiz (A, n = 255; B, n = 194; C, n = 197; D, n = 271; E, n = 296) followed by additional, more restrictive sample resizing. As the data represent the combination of nine quizzes, each 10 questions in length, given at the start of lessons over the course of a semester (i.e., nine quiz sittings spread over fourteen weeks), it is understandable that a certain degree of missingness would enter into the data due to tardiness and absenteeism. Though missingness was handled identically throughout the analyses, with '*' replacing vacant responses as a form of dummy placeholder (that commands Winsteps to ignore the item), no initial threshold restricted participant data strings. This meant that students who had answered all 90 items were an-

Table 1. Vocabulary quiz 1, 2, 3, 4, and 5 Item and Person Reliability and Separation Coefficients

Data Set	*N*-size	*k*-size	Item reliability	Item separation	Person reliability	Person separation
Vocabulary quiz A	188	87*	.96	5.18	.95	4.31
Vocabulary quiz B	185	87*	.97	5.74	.86	2.49
Vocabulary quiz C	173	88*	.97	5.52	.89	2.81
Vocabulary quiz D	233	89*	.98	6.80	.89	2.80
Vocabulary quiz E	162	84*	.96	4.98	.87	2.57

Note. Estimate coefficients represent Real RMSE rather than Model RMSE. *Values less than 90 indicate misfitting items dropped from analysis.

alysed side-by-side with those who had answered only 20 or 40 items. As missingness can affect the reliability and separation coefficients, as well as the misfit analysis, threshold restrictions were then applied, first at a 56% response rate (i.e., those who sat five of the nine quizzes) and then at the higher rate of 78% (i.e., those who sat seven of the nine). The goal of this iterative process was to attain the largest sample possible while maintaining acceptable estimates. (The n-sizes presented represent data from the 56% threshold.) The item estimates across all quizzes showed excellent reliability and either excellent or nearly excellent (in the case of quiz E) separation. For persons, the reliability and separation estimates were slightly lower and could be deemed as either good or acceptable. As the instruments under investigation fail to classify as *high stakes* assessments (i.e., used as gate-keeping devices such as in the case of certification or admissions procedures) that would require more stringent adherence to criteria standards, these coefficients represent acceptable values for further statistical analyses.

As mentioned, for this analysis sample sizes were kept as large as possible. Needlessly trimming performances can result in the removal of a unique portion of the sample, and, if absent, these participants' response strings could affect reliability and separation estimates. Often, estimates associated with person reliability and spread increase with the expansion of a sample size, as this growth often suffuses the data with greater variability. Smaller samples, especially drawn from a population of learners frequently affected by streamed educational instruction, result in relatively homogenous response patterns, where individuals display similar ability estimates that result in compressed separation and reliability coefficients. This is due to the fact that items one participant answers correctly other participants of an equal ability will also answer correctly. Conversely, the items one participant finds excessively difficult other participants of an equal ability will also find excessively difficult. Though classroom instructors often feel comfortable witnessing bell-shaped patterns in grade distributions stemming from classroom assessments, this shape expressed on a Wright map—revealed as clumps of persons aligned across from clumps of items—signifies either an inability in the item sample to distinguish between persons, or an inability in the persons sampled to differentiate between items. As sample sizes increase, so too does the potential for greater variability in person ability measures, especially at the lower and upper ends of the ability spectrum.

Item Statistics

Item statistics provide a means for analysing how closely the data fit model expectations and are reported as a mean-square value (MNSQ) and as a standardized value (ZSTD) for both persons and items. According to Linacre (2012), fit statistics such as these provide evidence of the amount of randomness or distortion created from unexpected response patterns. For MNSQ, the expected value is 1.0, which demonstrates either a very good fitting item or person. Values less than 1.0 represent over-predictive tendencies in the data. For items, excessively low values suggest insufficient sensitivity to distinguish between differing person abilities and are considered evidence of overfit. In contrast, MNSQ values greater than 1.0 signify underfit in the data, which is considered as "noise" resulting from excessive randomness. Underfit occurs when, for instance, a person with high ability (with a high probability of answering a greater number of questions correctly than a person with low

ability) responds incorrectly to an item modelled as easy; or, in contrast, when a low ability person responds correctly to an item of great difficulty (e.g., through cheating, guessing, or through particular knowledge exclusive to a subset of the items).

The infit statistic is a weighted standardized residual that is calculated using the expected score multiplied by the squared sum of the standardized residual for each person on each item. This reveals the variance for participants across all items. As the log-odd calculations are produced from expected performance of an individual across items, with the calculation considering both that person's prior response pattern as well as the item difficulty based on the response pattern of the sampled population, the model produces probabilistic assumptions about how each person should respond to each item. Items with a difficulty measure proximal to specific individual's level of ability carry greater sensitivity in terms of fit measures than items farther away. The expectation is that person ability measures will align with item difficulty estimates near those ability measures, and the more response patterns deviate from this assumption the greater the amount of noise enters the data through the fit indexes. In contrast, the outfit statistic is the average of the standardized residual and shows greater sensitivity to items with difficulty levels distant from the assumed ability of individual persons (Bond & Fox, 2007). Linacre (2007) states that a MNSQ range of 0.71 to 1.4 is considered very good, while a range from 0.77 to 1.3 is excellent. For the ZSTD, values greater than ± 1.9 represent misfit. McNamara (1996) suggested that for assessment instruments not deemed high stakes and that use dichotomous data from multiple-choice question formats, the range of 0.7 to 1.3 suffices.

The items presented in Table 2 follow the measure order produced for Vocabulary quiz A and match those provided in the Wright map from Figure 4, starting with the most difficult item (i.e., 18, conflicts) and extending down to the easiest item (i.e., 82, extinction). The Infit MNSQ values stretch from a high of 1.24 down to a low of .80 and reveal good fit. The Infit Z scores ranged from a high of 3.42 to a low of –2.95. The standard error of the model (SE) represents the precision of the measures, which is an indication of how well the persons and items fit model expectations. Here, higher values represent less precision and lower values more. As can be seen, the highest values are located on the extremes of the measurement scale, the top and bottom, which reflect items not equivalent to person ability measures for the sample. At the top, these items appear too difficult for the persons, and at the bottom this much larger cluster of items appears far too easy for the sample. This affects how well the model predicts these items to be reproducible across different samples. Items of this type, below and above the cluster of person abilities, fail to adequately target the ability levels of the sample.

Table 3 below shows the averages, maximum and minimum values, and the standard deviations for the standard error, infit MNSQ, infit *Z*, outfit MNSQ, outfit *Z*, and point measure correlation for the five data sets (Vocabulary quiz A, B, C, D, and E). For quizzes A, B, and E the infit MNSQ statistics showed extremely good fit, with quiz D revealing slight misfit and C a somewhat larger misfit. The standardized *Z* values, however, showed a degree of misfit.

As the data revealed a certain level of misfit, additional analyses were conducted to judge the level of misfit in regards to the potential effects this might pose to the statistical analysis. For this follow-up analysis, all misfitting items were removed from the data. For Vocabulary quiz A, this amounted to five items (44, 49, 53, 59, and 90); for Vocabulary quiz B, the total was four (5, 39,

Table 2. Vocabulary Quiz A Item Statistics Ordered by Measure

Item name	Measure	*SE*	Infit MNSQ	Infit *z*	Outfit MNSQ	Outfit *z*	P-M Cor
18 conflicts	74.2	1.90	0.98	0.08	0.85	−0.07	.19
60 standardized	68.8	1.30	0.91	−0.37	0.67	−0.95	.37
89 dissemination	64.3	0.90	0.88	−1.11	0.85	−0.79	.45
08 interaction	62.5	0.80	1.05	0.59	1.01	0.14	.32
76 precision	61.8	0.90	1.17	1.95	1.23	1.58	.21
48 demographic	60.9	0.80	1.00	0.07	0.91	−0.76	.41
17 constraint	60.3	0.80	0.97	−0.46	0.90	−0.97	.43
90 endorse	59.1	0.80	1.13	2.06	1.16	1.68	.25
39 explicit	58.8	0.90	0.91	−1.29	0.85	−1.50	.50
58 inference	58.7	0.80	0.95	−0.72	0.91	−0.91	.45
53 reliable	58.3	0.80	1.24	3.42	1.38	3.74	.16
69 polarization	58.1	0.80	0.94	−0.92	0.91	−1.05	.46
49 cleavage	57.3	0.80	0.87	−2.08	0.81	−2.24	.54
57 incident	57.3	0.80	0.92	−1.33	0.92	−0.84	.48
56 replicate	57.1	0.80	1.04	0.60	1.03	0.39	.37
50 agenda	56.8	0.80	1.01	0.23	0.99	−0.10	.40
37 emphasize	56.7	0.80	0.88	−1.87	0.82	−2.00	.53
70 collaboration	56.7	0.80	0.99	−0.15	0.95	−0.59	.42
47 diffusion	56.4	0.80	0.99	−0.06	0.98	−0.17	.41
87 Rigorous	56.3	0.80	1.02	0.35	1.07	0.90	.36
77 activate	55.8	0.80	0.90	−1.60	0.85	−1.76	.51
29 respondent	55.7	0.80	0.93	−1.11	0.91	−0.98	.47
68 elicit	55.6	0.80	0.93	−1.00	0.99	−0.05	.45
19 assess	55.5	0.80	1.07	1.11	1.13	1.46	.32
28 matrix	55.5	0.80	1.02	0.40	0.98	−0.21	.39
30 incorporate	55.0	0.80	0.98	−0.29	1.00	0.07	.41
59 predominantly	55.0	0.80	0.80	−2.95	0.75	−2.60	.59
67 placebo	54.6	0.80	0.94	−0.87	0.89	−1.00	.46
40 extract	54.3	0.90	0.94	−0.79	0.91	−0.83	.46
38 dependence	53.6	0.90	0.91	−1.13	0.88	−1.01	.48
79 viable	53.6	0.80	0.95	−0.65	0.96	−0.30	.43
36 median	53.1	0.90	1.09	1.06	1.10	0.82	.30
88 mutually	53.0	0.80	0.96	−0.49	0.91	−0.74	.42
26 optimal	52.8	0.80	0.87	−1.55	0.80	−1.63	.51
06 outcome	52.6	0.80	1.16	1.96	1.20	1.56	.23
80 subjective	52.4	0.90	1.01	0.13	0.99	−0.04	.37
44 randomly	52.3	0.90	1.23	2.42	1.30	1.96	.16
45 coherent	52.3	0.90	1.15	1.61	1.27	1.81	.22
16 exhibit	52.0	0.80	1.07	0.84	1.10	0.79	.30
46 entail	51.8	0.90	0.89	−1.14	0.85	−0.97	.47
78 sic	51.7	0.90	0.89	−1.24	0.78	−1.50	.49
71 postulate	51.5	0.90	0.99	−0.12	1.00	0.03	.38
10 cells	51.4	0.80	1.11	1.22	1.24	1.62	.24
35 compliance	50.6	1.00	1.02	0.20	1.02	0.17	.35
52 exploit	50.6	1.00	1.07	0.63	1.22	1.16	.27

Table 2. (Continued)

Item name	Measure	*SE*	Infit MNSQ	Infit *z*	Outfit MNSQ	Outfit *z*	P-M Cor
66 analogy	50.5	1.00	0.94	−0.51	0.78	−1.20	.43
07 construct	50.0	0.90	1.02	0.24	1.04	0.28	.32
04 issue	49.6	0.90	0.96	−0.32	0.99	0.01	.37
01 contrast	49.1	0.90	1.12	1.01	1.08	0.48	.23
27 orientation	49.1	1.00	0.96	−0.25	0.99	0.05	.37
32 allocation	48.8	1.10	0.92	−0.55	0.85	−0.61	.41
09 hypothesis	48.5	1.00	0.89	−0.85	0.73	−1.33	.44
83 parentheses	48.0	1.00	1.17	1.17	1.46	1.74	.09
42 justify	47.6	1.10	1.13	0.87	1.21	0.82	.19
86 absorb	47.0	1.10	1.06	0.38	1.31	1.12	.21
20 bias	46.7	1.10	1.01	0.08	1.11	0.47	.26
54 constituent	46.5	1.20	0.96	−0.17	0.92	−0.17	.32
55 align	46.5	1.20	1.01	0.10	0.96	−0.03	.29
43 patent	46.4	1.20	0.87	−0.72	0.68	−1.06	.42
03 feature	45.6	1.20	1.10	0.61	1.13	0.50	.17
84 ascribe	45.4	1.20	1.04	0.28	0.95	−0.03	.23
22 deviation	45.3	1.30	0.92	−0.32	0.92	−0.13	.33
64 compensate	45.2	1.30	0.96	−0.11	1.05	0.26	.27
61 propensity	44.4	1.40	1.05	0.29	0.94	−0.02	.21
02 definition	44.3	1.30	0.99	0.03	0.88	−0.22	.26
81 multivariate	44.3	1.40	1.08	0.43	1.11	0.40	.15
23 statistical	43.6	1.50	1.06	0.31	1.52	1.21	.12
14 attribute	43.4	1.40	0.93	−0.21	0.82	−0.35	.30
34 cohort	43.0	1.60	1.05	0.26	1.11	0.37	.18
73 resistant	42.5	1.60	1.13	0.53	3.81	3.68	−.05
51 locus	42.4	1.70	0.88	−0.29	0.46	−1.18	.36
41 heterogeneity	42.1	1.70	0.99	0.07	0.72	−0.44	.24
72 subtle	41.9	1.70	1.07	0.31	0.96	0.08	.15
85 periodic	41.8	1.70	0.86	−0.36	0.44	−1.28	.37
12 investigate	41.4	1.70	1.02	0.17	0.89	−0.06	.18
31 cognitive	40.9	1.90	1.07	0.32	1.07	0.32	.11
13 enhance	40.7	1.80	1.01	0.15	0.98	0.14	.16
21 spatial	40.5	1.90	1.04	0.22	1.63	1.10	.10
62 exogenous	39.9	2.10	1.03	0.20	0.87	0.01	.16
05 distribution	39.7	1.90	1.04	0.23	1.33	0.70	.09
25 fraction	39.6	2.10	0.94	−0.02	0.73	−0.25	.23
63 evolutionary	35.5	3.30	1.02	0.26	1.49	0.77	.05
24 restrict	35.3	3.30	0.93	0.11	0.21	−1.06	.25
11 perceive	34.7	3.30	1.02	0.26	0.84	0.10	.07
15 core	34.7	3.30	1.02	0.26	0.84	0.10	.07
33 genetic	32.4	4.60	1.00	0.33	0.49	−0.23	.10
75 asymmetry	32.0	4.60	0.94	0.27	0.15	−1.05	.20
65 discrimination	26.8	8.30	1.00	0.00	1.00	0.00	.00
82 extinction	26.4	8.30	1.00	0.00	1.00	0.00	.00

Note. SE = Standard error; MNSQ = Mean Square; *Z* = Standardized *Z*; P-MCor = Point-measure correlation

Table 3. Item Statistics Across Vocabulary A, B, C, D, and E Data Sets

Data set		*SE*	Infit MNSQ	Infit *Z*	Outfit MNSQ	Outfit *z*	Point-measure Correlation
Vocabulary quiz A (*k* = 89)	Mean	1.44	1.00	0.01	1.00	−0.03	.30
	Max	8.30	1.24	3.42	3.81	3.74	.59
	Min	0.80	0.80	−2.95	0.15	−2.60	−.05
	STDEV	1.30	0.09	0.97	0.39	1.11	.14
Vocabulary quiz B (*k* = 87)	Mean	1.26	1.00	0.01	0.97	−0.04	.28
	Max	4.60	1.27	4.60	1.68	3.98	.54
	Min	0.70	0.80	−3.49	0.31	−3.02	.00
	STDEV	0.77	0.09	1.03	0.25	1.07	.13
Vocabulary quiz C (*k* = 89)	Mean	1.31	1.00	−0.03	1.00	0.01	.32
	Max	8.30	1.55	3.98	1.98	5.32	.62
	Min	0.80	0.69	−3.50	0.19	−2.81	.00
	STDEV	0.98	0.12	1.21	0.29	1.25	.13
Vocabulary quiz D (*k* = 90)	Mean	1.32	1.00	−0.14	1.09	−0.01	.32
	Max	8.30	1.34	3.33	3.41	6.35	.67
	Min	0.70	0.73	−4.19	0.51	−3.53	−.13
	STDEV	1.20	0.11	1.24	0.51	1.54	.17
Vocabulary quiz E (*k* = 87)	Mean	1.64	1.00	−0.01	1.01	−0.01	.30
	Max	8.30	1.26	3.03	2.82	2.53	.59
	Min	0.80	0.80	−3.11	0.20	−2.79	−.03
	STDEV	1.55	0.09	0.96	0.40	1.09	.15

Note. M = mean; SD = standard deviation; * = median; + = interquartile range; ** = third quartile; ++ = first quartile; As the point biserial correlations, listed here as point measure correlations, are non-linear in nature, the mean and standard deviation must be viewed in reference to the extremes (Smith, Linacre, & Smith, 2003).

49, and 69); for C, it was ten items (3, 7, 9, 10, 24, 70, 77, 79, 80, and 88); for D, it was eight items (6, 10, 15, 56, 57, 64, 71, and 78); and for E, it was five (15, 47, 56, 79, and 87). (As the five vocabulary quizzes consisted of distinct sets of words, any duplication of item numbers above does not signify identical lexical items). Person measures with the misfitting items and new person measures produced without the misfitting items revealed little difference. For data sets showing normal distributions, Pearson product-moment correlation (r) was used, and for those showing slight non-normality the nonparametric Kendall rank correlation (τ) was used. For Vocabulary quizzes A, D, and E, the results of the Pearson product-moment analyses all showed extremely high correlations (r = .99); and for Vocabulary quizzes B and C, the results of the Kendall rank analysis showed high correlations as well (r = .95 and .90, respectively). As the data sets including the misfitting items represented no significant difference from those without the misfitting items, all items were retained.

Dimensionality

The final assumption of the Rasch model assessed in the preliminary analyses relates to instrument dimensionality, which has to do with whether or not the sum total of items on a particular

instrument have been designed to tap into a single underlying construct. For the five vocabulary quizzes, though half of the 90 items were multiple-choice (which arguably measures a receptive trait) and the other half fill-in-the-blank (which measures a productive trait), all items were designed to assess learners' awareness of academic vocabulary items drawn from the *Kyoto University Academic Vocabulary Database*. The meaningfulness of person and item estimates rests on whether or not the instrument functions in a unidimensional fashion. According to Wu and Adams (2007) "[l]atent variables are, in general, arbitrary constructs" (p. 22), in that they exist without a physical manifestation and are measurable only vicariously through items designed to tap into that particular trait or ability. As such, it is necessary to investigate whether particular items designed to measure a single construct reflect additional, unintended traits. Winsteps allows for the detection of whether an instrument designed to measure a single underlying construct does so effectively.

Confirming the unidimensionality of an instrument (here, five separate instruments) involves inspecting the raw variance explained by measures compared to the raw unexplained variance. The structure of the Rasch item residuals is presented in the form of a principle components analysis (PCA) of contrasts not explained by the measures. In the model, the primary dimension is hypothesized as the Rasch dimension, which is expected to account for the greatest degree of explained variance from the data. Beyond the first dimension, the assumption is that any additional contrast will consist of unexplained variance in the form of random noise. Linacre (2016b) states that if the variance explained by the items is greater than four times that found in the first contrast, this shows good item strength; and if the variance explained by the item measures is greater than 50%, this is considered extremely good. For Vocabulary quiz A, the explained variance amounts to 35.9%, which is well within the normal range. For the unexplained variance in the first contrast, if it measures at less than an eigenvalue of 3, it is thought to be good, and less than 1.5 is deemed excellent. In terms of percent, unexplained variance in the first contrast of less than 5% is considered excellent.

Table 4 below displays the amount of variance explained by the measures, persons, and items for Vocabulary quiz A. The raw variance explained empirically by the primary measure represented 35.9%, while the raw unexplained variance represented 64.1%. The breakdown of the percent that accounts for the measures consists of 12.4% for persons and 23.5% for items. Simply put, the items are explaining more than the persons, which mirrors prior reported results reflected in the item reliability and separation estimates. The best account of this is in the narrow performance spread displayed by the participant sample. The wider the spread, the greater the amount of variance explained by the persons, and therefore the measures.

In addition to the total variance explained by the measures, Table 4 also reveals the unexplained variance in contrasts one, two, and three. The first contrast accounted for an additional 2.5% of the variance, which is equivalent to 3.4 in eigenvalues or approximately 3 test items out of 90. The ratio of the first contrast (2.5%) to the measures explained (35.9%) is .07. These values are highly suggestive of unidimensionality, and the values for Vocabulary quizzes B, C, D, and E were nearly identical: .07, .07, .05, and .07 respectively. Winsteps also provides item loadings on additional contrasts, with threshold violations potentially indicative of additional contrasts. These loadings are calculated based on how participants systematically respond to items, which is then produced in a contrast

Table 4. Vocabulary quiz A Standardized Residual Variance

	Quiz A	
	Eigenvalue	**Empirical**
Total raw variance in observations	135.7	100.0%
Raw variance explained by measures	48.7	35.9%
Raw variance explained by persons	16.9	12.4%
Raw variance explained by items	31.9	23.5%
Total raw unexplained variance	87.0	64.1%
1st Contrast unexplained variance	3.4	2.5%
2nd Contrast unexplained variance	3.1	2.3%
3rd Contrast unexplained variance	2.5	1.9%

table, with each item given either a positive or a negative loading ranging from .0 to 1. Items that load beyond a ± .40 threshold can possibly reveal items or item clusters representative of additional constructs. When checked across the five vocabulary quizzes, these loadings registered items represented at the easiest and most difficult edges of the item difficulty hierarchy, rather than revealing distinctive patterns of lexical clusters. From this, there was little evidence to assume the presence of a secondary construct in any of the instruments.

Results

As the preliminary analysis revealed reliable measures for both item difficulty and person ability, the necessary requirements for further statistical analyses were met. The main analysis using the item and person measures in the form of transformed logits focuses on answering two basic questions. First, is there a significant difference in item difficulty due to question format? Second, is there a significant difference between learner performances from spring semester to fall semester?

Research Question 1: Relative Item Type Difficulty

The first set of analyses involved conducting an independent-samples *t* test on the item types, with the five vocabulary quizzes serving as distinct datasets. (Paired-samples *t* test could not be performed due to a mismatch in the item type totals, as misfitting items had been dropped from the analysis.) The goal of this analysis was to uncover whether the mean difficulty of the fill-in-the-blank questions was greater than that for the multiple-choice items. Each of the five datasets included learners from both spring and fall semesters combined.

An independent-samples *t* test was conducted to evaluate whether item type—multiple-choice and fill-in-the-blank—revealed different levels of difficulty in Vocabulary quiz A. The results indicated that the mean item difficulty measure for fill-in-the-blank items (n = 45, M = 55.75, SD = 5.22) was significantly greater than the mean difficulty measure for multiple-choice items (n = 42, M = 43.82, SD = 5.70), $t(83) = 10.16$, $p < .001$, with a Cohen's d reporting an effect size of 2.19. The mean difference between item type amounted to 11.93, with a 99% confidence interval showing 8.83 to 15.03.

For the remaining four vocabulary quizzes, the analyses showed similar results. For Vocabulary quiz B, the mean item difficulty measure for fill-in-the-blank items (n = 45, M = 56.16, SD = 4.95) was significantly greater than the mean difficulty measure for multiple-choice items (n = 42, M = 43.39, SD = 6.84), $t(74) = 9.91$, $p < .001$, with the mean difference between item type calculated as 12.77 and a 99% confidence interval showing 9.36 to 16.17. The Cohen's d showed an effect size of 2.15. For Vocabulary quiz C, the mean item difficulty measure for fill-in-the-blank items (n = 45, M = 55.67, SD = 4.43) was significantly greater than the mean difficulty measure for multiple-choice items (n = 43, M = 44.07, SD = 6.26), $t(75) = 10.00$, $p < .001$, with the mean difference between item type calculated as 11.60 and a 99% confidence interval showing 8.53 to 14.67. The Cohen's d reported an effect size of 2.15. For Vocabulary quiz D, the mean item difficulty measure for fill-in-the-blank items (n = 44, M = 57.24, SD = 5.19) was significantly greater than the mean difficulty measure for multiple-choice items (n = 44, M = 42.75, SD = 6.52), $t(82) = 11.53$, $p < .001$, with the mean difference between item type calculated as 14.49 and a 99% confidence interval showing 11.17 to 17.80 and a Cohen's d effect size of 2.46. Lastly, for Vocabulary quiz E, the mean item difficulty measure for fill-in-the-blank items (n = 42, M = 56.41, SD = 4.44) was significantly greater than the mean difficulty measure for multiple-choice items (n = 42, M = 43.58, SD = 6.90), $t(70) = 10.14$, $p < .001$, with the mean difference between item type calculated as 12.83 and a 99% confidence interval showing 9.48 to 16.18 and a Cohen's d effect size of 2.21. Plonsky and Oswald (2014), in a methodological review of effect sizes, place these effects sizes above the 90 percentile in reported size for articles in the field of applied linguistics. These results indicate that item type clearly affects item difficulty across all versions of the vocabulary quiz instruments, with receptive items showing lower levels of item difficulty than productive items.

Research Question 2: Person Measure Change

Data for this analysis represents responses across all 90 items provided by a sample of learners collected in the spring semester compared to item responses by a second sample of learners collected in the fall semester. (Vocabulary quiz A, B, C, D, and E all represent individual data samples here.) The item responses for the two groups for each data set are represented in the form of transformed ability measures produced as outputs from the Rasch analysis using Winsteps. The analysis method chosen was an independent samples t test.

An independent t test was conducted to evaluate whether person performances—fall semester compared to spring semester—revealed different levels of person ability over time. The results for Vocabulary quiz A indicated that the mean person ability measure for the second semester (n = 50, M = 58.54, SD = 4.46) showed a non-significant difference when compared to the mean person ability estimates for the first semester (n = 50, M = 57.04, SD = 4.45), $t(98) = 1.68$, $p = 0.10$, with the mean difference between person performances for the two semesters calculated as only 1.50 [−0.84, 3.84] greater over time. The reported Cohen's d effect size was 0.34.

For the remaining four vocabulary quizzes, the analyses showed mixed results. For Vocabulary quiz B, the mean person ability measure for the second semester (n = 50, M = 56.35, SD = 4.66) also showed a non-significant difference when compared to the mean person ability estimates for the first

semester (n = 50, M = 56.24, SD = 3.32), t(88) = 0.13, p = 0.89, with the mean difference between person performances for the two semesters calculated as a 0.11 [−1.50, 1.72] difference between the two samples. The reported Cohen's d effect size was 0.03. However, for Vocabulary quiz C, the mean person ability measure for the fall semester (n = 59, M = 57.01, SD = 4.23) showed a significant difference when compared to the mean person ability estimates for the spring semester (n = 59, M = 54.39, SD = 3.89), t(115.20) = 3.50, p < 0.01, with the mean difference between person performances for the two semesters calculated as 2.62 [1.14, 4.10]. The reported Cohen's d effect size for this analysis was 0.64. In contrast, just as with quiz A, for Vocabulary quiz D, the mean person ability measure for the fall semester (n = 35, M = 55.55, SD = 4.34) showed a non-significant difference when compared to the mean person ability estimates for the spring semester (n = 37, M = 55.05, SD = 4.96), t(70) = 0.45, p = 0.65, with the mean difference between person performances for the two semesters calculated as 0.50 [−1.69, 2.69]. Here, the effect size was 0.11. Lastly, for Vocabulary quiz E, the mean person ability measure for the fall semester (n = 69, M = 57.57, SD = 4.71) showed a significant difference when compared to the mean person ability estimates for the spring semester (n = 69, M = 55.48, SD = 3.88), t(131.00) = 2.85, p < 0.01, with the mean difference between person performances for the two semesters calculated as 2.09 [0.64, 3.55]. The effect size for this was 0.48. These results suggest a differential learning effect across semesters, with the potential for particular subgroups of the first-year cohort improving over time and others remaining static.

Discussion

Results from the item analysis revealed that the five vocabulary quizzes were reliable instruments for measuring first-year learner awareness on a range of vocabulary items designed for a set of academic writing and listening courses offered to first-year learners across a variety of departments. Since the vocabulary quizzes have been determined to be internally consistent, use of these items with some revisions can be considered for assessing future cohorts. To further improve test reliability, a test-retest reliability measure could be performed by having the same group of learners repeat selected items from various quizzes in the spring and fall semester.

Item Difficulty: Productive vs. Receptive

The first research question involved the relative difficulty of the two item types. Results from this analysis support prior research (Laufer, 2005; Webb, 2008) that learners' receptive knowledge of vocabulary is greater than their productive knowledge. The majority of fill-in-the-blank items (those requiring productive knowledge) were above the mean level of difficulty, and most items appearing below the mean were multiple-choice questions (those requiring receptive vocabulary knowledge). This separation can be seen in Figure 4 (Quiz A) where 41 of the 47 items above the mean of are fill-in-the-blank items (87%), while 38 of the 42 items below the mean level of difficulty are multiple-choice items (90%). The remaining quizzes revealed identical ratios of item types above and below the mean difficulty level. These results may also echo research suggesting that there are degrees to vocabulary knowledge and that learners only acquire full knowledge of individual lexical

items when they are able to use the target vocabulary in writing or speech (Laufer, 1997), although each of these tasks would require different dimensions of vocabulary knowledge (i.e., orthographic knowledge in the case of writing, and phonetic knowledge for speech). The assumption, therefore, is that learners should be able to recognize words in a multiple-choice question that were previously answered correctly in the fill-in-the-blank question.

Yet the dichotomous division of items in terms of difficulty may not fully explain the learners' depth of knowledge of the target vocabulary. Receptive knowledge and productive knowledge are different degrees of knowledge from that which is required to recognize or recall either the meaning or form. For example, learners in this study may be able to select the meaning of the L2 target word from a list of distractors, but to recall the meaning of the L2 word and produce the L1 translation may present a greater challenge. Laufer and Goldstein (2004) argue that these four degrees of knowledge—receptive/productive and recall/recognize—can be organized into hierarchical modalities, with receptive recognition the least difficult, followed by productive recognition, receptive recall, and finally productive recall the most challenging. The hierarchy is not frequency dependent. Thus, at every level of frequency, including specialized academic word lists, the same hierarchy of difficulty in vocabulary knowledge is posited to exist. The question type in the vocabulary quizzes targeted only the bottom (multiple-choice question) and top (fill-in-the-blank question) of this hierarchy. By adding a variety of questions to reflect these different modalities, items would be able to measure subtler differences in vocabulary knowledge (Milton, 2009; Schmitt, 2010). The goal, therefore, would be to investigate what additional question types might provide greater description of learner knowledge across a range of academic vocabulary.

In addition to question type as a factor to explain the clear division of items by level of difficulty, the low-frequency of the academic vocabulary might also have produced these results. The *Kyoto University Academic Vocabulary Database* was designed to cater to the needs of students in this course and facilitate academic reading, research, and writing. As previously explained, the list of target vocabulary contains words that were considered frequent across academic disciplines but that are not high-frequency compared to non-academic word lists. Despite the learners having a number of years of English study, materials utilized in previous English courses at the junior and high school level most likely did not include these low-frequency words. Comparative studies have noted a decline in vocabulary types and tokens in senior high school textbooks (Chujo, Nishigaki, Hasegawa, & Uchiyama, 2008), while research investigating vocabulary characteristics and readability in junior high school textbooks revealed that there is no correlation between vocabulary contained in such textbooks and commonly used frequency lists such as the GSL and AWL (Kitao & Tanaka, 2009). Thus, the low exposure to the target vocabulary prior to encountering the 477 words may have limited the learning of these words to the form-meaning dimension.

As Schmitt (2008) explains "[w]hile it is true that the form–meaning link is the first and most essential lexical aspect which must be acquired, and may be adequate to allow recognition, a learner needs to know much more about lexical items, particularly if they are to be used productively" (p. 333). Results from this research show that learners may have achieved this first step in acquiring the academic vocabulary through explicit study. Laufer (2005) argues, "if learners want to reach

active knowledge of new words, or activate words already known passively, they will have to focus repeatedly on these words and practice them in demanding tasks" (p. 233). The question remains whether a more systematic approach through materials development can provide learners with adequate opportunity to familiarize themselves with low-frequency academic words.

Based in part on this and extant research, an updated version of online listening materials for the EWL course are currently being constructed. The materials themselves are being designed to link more explicitly with the vocabulary items. The vocabulary practice sections are being rewritten to better introduce the target vocabulary through L1-L2 translations and example sentences, as well as to better scaffold vocabulary knowledge through a series of multiple-choice, true-false and cloze activities. The ultimate goal is to increase the rate at which learners encounter the target vocabulary through customised academic listening lectures and conversations.

Learner Performance

One key area of educational assessment at the curricular level is measuring the cumulative development of cohorts as they progress through set stages, steps, or sequences in a unified curriculum. This starts with designing appropriate means of analysing learner growth over time so as to ensure that the materials presented to learners amounts to the best option available in regards to teachability, learnability, and practicality. Though most curriculum designers conceive of materials as learnable elements within an organized progression, portions of any curriculum will always be more easily learned than others, and portions of learner cohorts will face difficulties while other portions will not. When evidence of limited learning occurs, it is crucial that language educators uncover weaknesses in the instructional method or content and devise alternative approaches to introducing troublesome material. In terms of the academic list designed as part of the EWL course, learners in both spring and fall semesters encountered the same range of vocabulary items (i.e., 50 words a week provided in a set sequence—1–50, 51–100, 101–150 up to 477, with the final assessment including a total of 77 potential items).

The mixed results from the person analysis require some degree of sceptical review, as the population samples for spring and fall semesters included distinct subgroupings of first-year learners. Only two of the five *t* tests showed significant differences across semesters, and this may have been due to patterns involved in sample selection relating to issues in course scheduling, quiz implementation, and data collection. In addition, only a limited number of teachers teaching the courses participated in data collection, with no controls placed on gathering year-long data for the same students. As a result, this meant that learning gains could only be measured in the form of a cross-sectional design. Therefore, despite the homogenous sample population, caution should be taken in interpreting changes in performance over one academic year.

The learners' own time-outcome investment may also have accounted for what may appear to be low learning gains. Conditions in the university curriculum in terms of time and workload call for students to manage time efficiently. Factors that may influence time management decisions include the number of hours required to take department-specific classes as well as the greater challenges of the EWL course itself—both in the type and amount of listening as well as the type and amount

of writing work, each of which carried a greater weight in respect to the allotted course grade than the vocabulary portion (i.e., writing amounted to 60% in the spring and 50% in the fall; listening amounted to 30% in the spring and 40% in the fall; whereas, vocabulary amounted to 10% in both spring and fall). The point value of each vocabulary item in relation to the total percent allotted to the vocabulary portion of the course grade represented 0.11% and to the entire course grade just 0.01%. With a grade value this low, combined with the relative amount of time necessary to become competent in using extremely infrequent vocabulary items within constrained sentence patterns (i.e., fill-in-the-blank question types), learners might have considered their time better spent in other pursuits than memorizing lexical items they might seldom encounter and potentially never use. The point here is that changes in the curriculum for future cohorts of learners is now underway, with a goal of enhancing learner experience and contact with each of the tested vocabulary. The purpose of this is to make opportunities available for learners to repeatedly encounter these words in use across contexts and modalities.

Limitations

There are several methodological limitations with this study. First, due to the nature of independent person samples, it was not possible to adequately measure growth over time. While this was not necessarily the intent of the study, the approach and the methods used could be improved in the future. Second, as expected when trying to evaluate the efficacy of instructional materials within an operating, semi-structured curriculum, convenience sampling meant that control of the sampling for the study was not possible. Finally, the lack of even a truncated pre-test assessment of participants' prior knowledge makes gauging proficiency and performance considerably more fraught.

Future Research

In response to some of the limitations outlined above, several improvements to future studies would include a pretest/posttest design to gauge semester-to-semester learning attributable to individual learners. In addition, by working in conjunction with faculty who teach the same group of students either in the spring or fall semester, data could be collected for a single set of students over two semesters, and learning gains over time could be more accurately measured.

This study revealed that the level of difficulty of an item utilized in the vocabulary quizzes depended on it being receptive or productive. However, there is much about the item difficulty that has not been accounted for. First, the item format may not be equal for comparison of productive and receptive knowledge (Webb, 2008). The difference in cognitive load between recognising the correct word in answering the multiple-choice question and recalling the word in the case of a fill-in-the-blank item may not lend to an accurate assessment of receptive and productive knowledge. It is unclear whether the learners would be able to answer a reversed question type, from fill-in-the-blank to multiple-choice, correctly due to question type factors, for example misleading distractors in the case of multiple-choice questions or overly explicative contextual cues in the fill-in-the-blank ques-

tions. Therefore, future research should include items with reversed question type from receptive to productive and vice versa to understand whether results truly reflect the learners' knowledge of the target vocabulary rather than the level of difficulty of the question type.

Another aspect of the quiz items to account for in future studies is lexical difficulty as measured by frequency. It is not yet clear what makes an item more or less difficult within the same question type. For example, the relative lexical difficulty of the distractors in multiple-choice questions and fill-in-the-blank prompts should be considered to understand whether there is a particular threshold that might affect item difficulty. Moreover, measuring item difficulty within the same pool of item types by looking at the frequency range of the specialised academic word list could uncover whether or not relative lexical infrequency affects item difficulty. Other potential vocabulary and item-dependent variables to include in this type of analysis are loanwords, number of letters, item-stem complexity, and distractor complexity. Understanding which of these variables has the greatest effect on item difficulty could help both in the revision of items for assessment and in the creation of new materials. Moreover, looking beyond the lexical-grammatical effects on item performance measures, accounting for other variables such as gender or area of specialisation (i.e., faculty and department) could determine whether any of the items are biased in favour of a subsection of the respondents.

Results from this study and future analyses of the data collected can bring to light problematic areas in the explicit study of academic vocabulary and assessment of the target vocabulary. To complete this picture, subsequent research should include the collection of qualitative data in the form of surveys about time spent studying, use of the wordlist, and other materials or tools used in preparation for the vocabulary quizzes. In addition to accounting for the learners' use of study strategies, by observing the interaction between test-takers and test items in a think-aloud protocol, more could be understood about the learners' use of test-taking or coping strategies.

Conclusion

This research has provided an in-depth investigation of the validity of a set of vocabulary assessments designed to measure learner knowledge of a set of 477 pre-determined academic vocabulary items. Results show that the vocabulary instruments, in the shape of nine quizzes constructed in five versions (i.e., 45 total quizzes), revealed good model fit based on a Rasch analysis of the combined items for each quiz version. Follow-up statistical analyses in the form of *t* tests revealed information about both item type and potential learner gains over time. The item-type analysis revealed that items measuring learner receptive knowledge proved easier than those measuring productive ability. The person-level analysis of gains over time showed somewhat mixed results, with only two of the five learner populations revealing gains from spring semester to fall. This analysis was performed as part of a larger needs analysis project with results providing insights into how best to structure additional in-house created materials.

References

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocab-

ulary tests. *Language Testing*, *16*(2), 131–162.

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

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*, 653–675.

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list. Retrieved from http://www.newgeneralservicelist.org.

Chujo, K., Nishigaki, C., Hasegawa, S., & Uchiyama, M. (2008). The impact of *yutori kyouiku*: A comparative study of 1988 and 2006 high school textbook vocabulary. *English Corpus Studies*, *115*, 57–80.

Chung, T., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, *15*(2), 102–116.

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

Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252–267). Cambridge: Cambridge University Press.

Farrell, P. (1990). Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary. (CLCS Occasional Paper No. 25). Dublin, Ireland: Trinity College, Centre for Language and Communication Studies.

Fisher, W. P. Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095.

Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly*, *51*(4), 844–867.

Kitao, K., & Tanaka, S. (2009). Characteristics of Japanese junior high school English textbooks: From the viewpoint of vocabulary and readability. *Journal of Culture and Information Science*, *4*(1), 1–10.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal*, *73*(4), 440–464.

Kyoto University EAP Vocabulary Research Group and Kenkyusha (2009). *Kyodai gakujutsugoi deitabeisu kihon eitango 1110*. [Kyoto University Academic Vocabulary Database]. Tokyo: Kenkyusha.

Laufer, B. (1997). What's in a word that makes it hard or easy: Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140–155). Cambridge: Cambridge University Press.

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, 5, 223–250.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51.

Leki, I., & Carson, J. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly*, *28*(1), 81–101.

Liddicoat, A. (2007). *An introduction to conversation analysis*. *Lo*ndon: Continuum.

Linacre, J. M. (2012). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2016a). Winsteps® (Version 3.92.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2016. Available from http://www.winsteps.com/

Linacre, J. M. (2016b). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com

Martin, A. (1976). Teaching academic vocabulary to foreign graduate students. *TESOL Quarterly*, *10*(1), 91–97.

McNamara, T. (1996). *Measuring second language performance*. Harlow, Essex, UK: Addison Wesley Longman Ltd.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Mochizuki, M., Aizawa, K., & Tono, Y. (2003). *Eigo goi no shido manyuaru*. [Teaching manual of English vocabulary]. Tokyo: Taisyukan shoten.

Mondria, J. A., & Mondria-De Vries, S. (1994). Efficiently memorizing words with the help of word cards and "hand computer": Theory and applications. *System*, *22*(1), 47–57.

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, P. (1997). Vocabulary size, text coverage and word lists. In N. M. Schmitt, M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge: Cambridge University Press.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2010). *Researching and analysing vocabulary*. Boston: Heinle ELT.

Plonsky, L., & Oswald, F. L. (2014). Methodological review article: How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

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

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Tajino, A., Stewart, T., & Dalsky, D. (2010). *Writing for academic purposes*. Tokyo: Hituzi Syobo Publishing.

Tajino, A., Dalsky, D., & Sasao, Y. (2009). Academic vocabulary reconsidered: An EAP curriculum-design perspective. *Journal of Teaching English as a Foreign Language and Literature*, *1*(4), 3–21.

Schmitt, N., & McCarthy, M. (Eds.). (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. London: Palgrave MacMillan.

Smith, R. M., Linacre, J. M., & Smith, Jr., E. V. (2003). Guidelines for manuscripts. *Journal of Applied Measurement*, *4*, 198–204.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95.

West, M. (1953). *A general service list of English words*. London: Longmans, Green and Co.

Wilkins, D. A. (1972). *Linguistics and language teaching*. London: Edward Arnold.

Wu, M., Adams, R., & Educational Measurement Solutions (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne, Victoria: Educational Measurement Solutions.