# BY THE NUMBERS: THE RATIONALE FOR RASCH ANALYSIS IN PLACEMENT TESTING

MARTYN CLARK

*University of Hawai‘i at Manoa*

## ABSTRACT

Placement tests are usually designed to assess relative language ability within the range of a particular program. Test scores are generally interpreted as measures of language ability, and students are compared and placed in accordance to them. This paper argues that an application of the Rasch model to placement situations is not only warranted by the assumptions of the placement process, but also that great benefits can be achieved by examining items and persons that do not fit the Rasch model. To illustrate these points, the University of Hawai‘i English Language Institute Academic Listening Test is analyzed and discussed.

This paper uses a Rasch analysis perspective to examine the Academic Listening Test (ALT) used by the English Language Institute (ELI) at the University of Hawaii at Manoa for placement into academic listening courses. This is not a validation study of a new test but rather a reevaluation, from a Rasch measurement perspective, of a test that has been used for almost a decade. Although this is certainly not the first application of a latent trait approach to placement test analysis (e.g., Blais & Laurier, 1995; Kondo-Brown & Brown, 2000; Sasaki, 1991), the diagnostic information available with the Rasch approach is rarely exploited. This reevaluation is the first step in revising and updating the ALT.

### *Placement Testing and Test Scores*

Placement testing in language programs is primarily concerned with assessing students' language proficiency for the purpose creating relatively homogeneous groups for instructional purposes (Bachman & Palmer, 1996; Brown, 1996). For a language school, the placement test may cover a wide range of ability and subsequent placements may range from beginning to advanced instructional classes. In other cases, such as support language programs for international students studying in the US, this placement decision may also include the determination that the student in question does not need further language instruction because he or she has exceeded the level of instruction provided by the service program. This is often the case in the university setting in which students for whom the language of instruction is not their native language must demonstrate a minimum level of language for conditional admittance into the university and a second, higher level of proficiency to take a full course load. In this type of situation, the placement decisions are usually not over the whole range of language proficiency, but rather within a relatively narrow band of language ability, specifically between the admittance level and the exemption level (Brown, 1989).

Placement test scores are interpreted as measures of language ability, that is, a higher score on the test indicates a higher level of language ability and thus warrants placement in a more advanced language course. Of course, the actual designations of courses as being at a certain ability level (i.e., *Intermediate Listening*) tend to be arbitrary and program-specific.[1] Given the same number of students and range of language abilities, one program may have the resources to offer small classes representing fine distinctions in ability whereas another program may merely divide the group into *beginning* and *advanced* classes. Regardless of actual placement procedures employed, the assumption still remains that the placement test is distributing students along a continuum of language abilities from which instructional groupings can be created. In fact, perhaps it is more appropriate to say that the logic of placement testing as it is usually carried out requires that the placement instruments distribute students along a continuum of language abilities in the domain of interest.

[1] An exception to this general rule would be programs which tie their courses to common proficiency rating scales (e.g., ACTFL, ASLPR) but even here differences in program resources can lead to classes composed of students from wider or narrower chunks of the scale.

### *The Current Study*

This paper starts from the premise hinted above, namely that the logic of placement testing, at least as it is carried out by this program and probably many other similar programs, makes the implicit assumption that the total placement test score of a student is a sufficient indicator of language abilities and thus can be directly compared across students for the purposes of placement. Therefore, if the ALT is to be useful as a measurement instrument, it should have the following characteristics: (a) a higher score on the test represents a greater level of listening ability, (b) the items are targeted to the population that the test is designed for, and (c) the items do not function differentially for subgroups of examinees. This paper will start with a description of the ALT, then outline the salient points of the Rasch model and the rationale for analyzing the test from this perspective. Next, data used in this study will be described and the results of the analysis will be presented with reference to the necessary characteristics cited above. The final section summarizes the points made in the paper.

## BACKGROUND AND RATIONALE

### *The Academic Listening Test (ALT)*

The ALT is one of a battery of tests used to determine if newly admitted students for whom English is not a first language have sufficient English ability to take a full load of credit bearing courses with no additional language support. This test is only administered to students who have not provided evidence of sufficient language proficiency (a particular number of transferable credits from English-medium institution, scores above set criteria on standardized tests, etc.) at the time of registration for classes. Thus the range of language ability to be tested is rather narrow as students with very low ability would have been denied admission to the University outright and students with high ability have already been exempted from additional language study. Three placement decisions are possible based on the ALT score: (a) additional study at the intermediate level, (b) additional study at the advanced level, or (c) exemption from further study.

The ALT consists of four sections, the breakdown of which is shown in Table 1. All of the items use the multiple-choice format. This format was chosen primarily for practical reasons as the results of the test must be made available as soon as possible so students can continue with the registration process and the multiple-choice format allows for the machine scoring of tests.

Table 1
*Overview of the Academic Listening Test (ALT)*

| Section | Task | Topic | Item numbers |
|---|---|---|---|
| Section One | Listen to a short passage and answer questions pertaining to the content | Volcanic origins of Hawaiian islands | 1 – 4 |
| | | Development of motion pictures | 5 – 10 |
| | | MLV Train | 11 – 14 |
| | | Missing library book | 15 – 20 |
| Section Two | Determine the meaning of a word after hearing it used in a sentence | Vocabulary | 21 – 24 |
| Section Three | Listen to two sentences and determine the best word or phrase to connect them | Transitions | 25 – 29 |
| Section Four | Listen and take notes for a ten minute lecture then answer questions pertaining to that lecture | Lecture – Culture and language | 30 – 40 |

The conception of language ability underlying the ALT is essentially one of communicative competence. Though there have been revisions and reformulations in the literature (e.g., Bachman, 1990; Bachman & Palmer, 1982; Canale, 1983; Canale & Swain, 1980), the essential idea is that the ability to communicate in a language entails not only the ability to manipulate the formal structure of the language properly (organizational competence), but also the ability to produce discourse that is appropriate for situation and context (pragmatic competence). These competencies are in turn composed of sub-competencies at finer levels of scale such that, for example, organizational competence entails grammatical competence (the formulation of grammatically appropriate sentences) and discourse competence (the arrangement of a series of grammatical sentences into a larger chunk of discourse, such as a lecture). It is assumed that gradual increases in the various sub-competencies eventually manifest themselves as an increase in overall competence. This is true of comprehension as well as production. As one becomes more proficient, all other things being equal, one is better able to handle more demanding tasks. To give an example, the ability to distinguish between the language one is studying and another language might be considered an easy task whereas the ability to listen to a lecture on an unfamiliar topic and recount the main points would be a task that requires considerably more listening ability (Nunan, 1989). Of course, in the case of listening, the characteristics of the stimulus itself can impose greater or lesser demands on the listener even though the listening task is similar. In other words, a lecture composed of easy lexis delivered clearly with prosodic emphasis on the main points would be considerably easier than a lecture of identical length composed of uncommon words delivered in a mumbling monotone voice at breakneck speed.

In terms of word choice and discourse style, the language on the test is academic in nature. Sections One and Four represent the essential listening task of attending to a spoken message for content. The topics were chosen for their interest and generality. The language in these passages reflects what might be considered general academic language such as would be found in an introductory course. It is assumed that someone with greater listening ability would have more success at comprehending the passages. In all cases, the passages were recorded using a script, so they are artificial in the sense that they do not contain the false-starts and self-corrections that might be produced by someone speaking extemporaneously. The exception to this is the lecture in Section Four which, though scripted, was intentionally recorded in a more relaxed manner and is more akin to someone lecturing from notes than reading a prepared manuscript, complete with false starts, hesitations, and fillers. Sections Two and Three reflect an interest in the sub-components of language ability discussed above, namely the ability to infer unknown words from context (Section Two) and the ability to recognize appropriate discourse structuring devices (Section Three). It has been noted that confusion can arise when dealing with components which are hierarchical in nature if one is not cognizant of the appropriate level of scale that should be considered for the measurement purpose at hand (Andrich, 2002a, 2002b). A substantive question for the analysis of the ALT is whether these four sections represent the same level of scale or not.

### ***The Case for the Rasch Model***

Dunkel, Henning, and Chaudron have argued that "unless some implicational or Guttman-type scale can be formed with monotonic increment of person ability and task difficulty in the same response matrix, whatever we choose to label listening comprehension would not qualify as a unitary measurement construct, and the reporting of unitary scores as a reflection of comparative performance would be misleading" (1993, p. 182). The Rasch model is ideally suited to this task for two reasons. First, both items and persons are on the same metric and, second, the total score is a sufficient statistic (Linacre, 1992; van der Linden, 1992). This means that, provided the data fit the model, the total score contains all of the information about an examinee's ability and thus, "the classification of persons according to their total scores is justified" (Andrich, 1988b, p. 38). In the Rasch model, the probability of a person succeeding on a given item is dependent upon the ability of the person and the difficulty of the item. The more able a person is in relation to a given item, the greater probability there is of that person being successful on that item. Unlike the Guttman model, the Rasch model is probabilistic rather than deterministic and recognizes that the same total score can be arrived at by different combinations of items, with the Guttman structure being the most probable pattern (Andrich, 1985).

The model has also been described as axiomatic (Bond & Fox, 2001; Wright, 1997) in the sense that, as a mathematical model, it requires data which represent a unitary construct in accordance with a particular theory of that construct (Andrich, 1989). This does not mean that the construct cannot have several psychological dimensions or components. Borrowing an argument and analogy from Thurstone (1928), it is impossible to represent the entire complexity of an object as a single value; even something as concrete as a table (p. 215). There is always a certain loss of information in any measurement. That is, it is impossible to measure a table without specifying what aspect of the table (weight, height, etc.) will be measured. Taking the analogy further, even though most people would agree that it is perfectly acceptable to talk about the weight of tables and make comparisons between them on that basis, this does not imply that a table's weight is readily and consistently determinable from its constituent parts or properties. Certainly, tables of the same weight can differ in terms of color, size, material used in construction, style, degree of wear, number of missing parts, etc., with many of those factors contributing directly to the overall weight. Nevertheless, even though there are many factors which contribute to the weight of the table, it is not required that they all be specified or even present in consistent proportions for a useful measurement of weight to take place.

In terms of the ALT, this means that describing listening comprehension in terms of a single test score is not inherently incompatible with the notion that there are many dimensions to listening ability, and, that students may differ along some or all of those dimensions. The important idea is that it is possible to conceptualize students along a continuum of listening ability in which different locations on the continuum correspond to the notion of having "more" or "less" listening ability. Items on the test are assumed to represent tasks which require relatively more or less listening ability and to the extent that the items on the test function together to define that continuum, the model holds and the total score is a sufficient measure of this ability.

### *The Model of Analysis*

For this particular analysis, Andrich's extended logistic model (Andrich, 1988a) will be employed. This model is an extension of Rasch's simple logistic model (Rasch, 1960/80) to accommodate ordered response categories. This model is called for in this instance because the nature of the listening test is such that several questions are associated with any given listening passage (cf. Table 1). Because of this, it is prudent to treat the sum of items for a given passage as a whole rather than as individual items as the dependencies between items associated with a passage are not likely to be the same as the dependencies between items across passages (Andrich, 1988a). For each passage, a potential score of $k_m$ is possible, where $k_m$ represents the total of all items associated with that passage. The movement from a score of x = 0 (no items correct) to a score of x = $k_m$ (all items correct) can be envisioned as progressing through a series of transition points ($\tau_1$, $\tau_2$, etc.), namely those points at which the probability of getting one more item correct (the point between x + 1 and x + 2, etc.) is exactly as likely as getting one more item incorrect. It should be noted that x + 1 does not refer to a specific item in the group of questions, but merely an increase in the score and different persons will achieve a score of x + 1 through different combinations of items. (It should also be noted that the model for dichotomous responses is simply a special case of the extended model in which $k_m$ = x, that is, the only transition point is between a score of 0 (incorrect) or 1 (correct).) The extended logistic model can be written as follows:

$$\Pr\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp(\kappa_{xi} + x(\beta_n - \delta_i))$$

$$\text{Where: } \gamma_{ni} = \sum_{k=0}^{m} \exp\left[ -\left( \sum_{x=0}^{k} \tau_x \right) + k(\beta - \delta) \right]$$

This model is analogous to Master's Partial Credit Model (Masters, 1988; Wright & Masters, 1982) and Andrich's Rating Scale Model (Andrich, 1978), the only substantive difference being whether threshold values are treated as variable for different items (partial credit) or consistent across items (rating scale). In the case of the ALT, items in Sections Two and Three are treated as having two ordered categories (i.e, dichotomous – a score of 0 for incorrect and 1 for correct) while items in Sections One and Four are treated as having multiple ordered categories with the number of categories dependent on the number of items associated with each passage. Thus, a reading passage with four associated items such as the MLV passage is treated as having five ordered categories (scores of 0 through 4 correct).

## METHOD

### *Participants*

The data selected for this project are from the previous five semester administrations, comprising two and a half academic years. By definition, all of the students taking the ALT have demonstrated sufficient English proficiency to be accepted to the University of Hawaii but not sufficient to be exempt from further language instruction. The total number of students was 692. Students were classified according to L1 background and academic status. Though more than 40 first languages were reported (see Appendix A for a complete listing), Chinese, Japanese, and Korean speakers were much more numerous than others. For this reason, only those three languages were explicitly coded, other languages were classified as *Other*. Because students were inconsistent in specifying the dialect of Chinese spoken, both Mandarin and Cantonese were coded as *Chinese*. Under this classification, 137 students were coded as *Chinese*, 283 as *Japanese*, 120 as *Korean*, and 152 as *Other*. For academic status, a simple graduate/undergraduate distinction was used as it was felt that this would most easily differentiate, albeit roughly, between those students with relatively more and relatively less college experience. True undergraduates as well as exchange students at the undergraduate level were coded as *undergraduate* (*n* = 458). The *graduate* category (*n* = 234) includes graduate students, exchange students at the graduate level, and students seeking a second bachelor's degree. Though the ALT contains language that could be considered academic, the assumption is that a students' academic readiness to attend the university has been ascertained from their high school or college record. Therefore, the ALT should be primarily a measure of language ability, albeit with an academic focus, rather than an additional measure of academic preparation, which might favor students with previous college experience.

### *Materials*

The data used in this study are from scheduled administrations of the ALT. The ALT is part of a language test battery that also includes reading and writing tests. To allow some flexibility in scheduling and to accommodate the number of students that must be tested (especially in the fall which sees the greatest number of incoming students), the ALT is administered several times, usually three days apart, just prior to the start of each semester. Students are free to sign up for whichever administration best suits their schedule. The entire placement test battery takes approximately four hours and the ALT is the third test in the sequence.

The ALT is administered under controlled conditions by members of the ELI teaching staff in a large auditorium. Examinees sit in rows and are separated from each other by at least one empty seat. Each examinee receives a test booklet containing instructions and questions and a machine-readable answer sheet, both of which are collected immediately following the test. All of the instructions and passages for the ALT are delivered via audio recording. Students are not allowed to consult dictionaries or take notes during the test. The only exception is for the last section of the test in which the students must listen to a ten-minute lecture. Immediately prior to this section, students are given two sheets of blank paper on which to take notes. This paper is also collected at the end of the test for security reasons, but the notes themselves are not scored.

### *Analysis*

Although the data were collected at various times (i.e., in various administrations), Table 2 shows that, for the most part, the ALT produced similar means, score ranges, and standard deviations. It should be noted here that Table 2 represents the descriptive statistics from each actual test administration. Although a total of 804 students took the test during this period, not all of the records were used in the combined analysis for this paper.  Records were excluded for one of two reasons: (a) the student had taken the test more than once, in which case the retest score was excluded or (b) the status or language background information was not available for the student in question. This second reason was the most frequent as this information is provided on a voluntary basis and is not required or used for placement purposes. In no instance was test performance used to exclude students.

Table 2

*Descriptive Statistics for the ALT*

| | Fall 2001 | Spring 2002 | Fall 2002 | Spring 2003 | Fall 2003 | Combined[a] |
|---|---|---|---|---|---|---|
| *N* | 229.00 | 92.00 | 213.00 | 66.00 | 204.00 | 692.00 |
| *k* | 40.00 | 40.00 | 40.00 | 40.00 | 40.00 | 40.00 |
| Mean | 23.38 | 24.00 | 24.11 | 23.20 | 23.48 | 22.94 |
| Median | 23.06 | 24.60 | 23.71 | 23.50 | 23.41 | 23.00 |
| *S* | 6.05 | 5.81 | 6.14 | 6.81 | 6.01 | 5.75 |
| High | 39.00 | 39.00 | 40.00 | 36.00 | 37.00 | 40.00 |
| Low | 7.00 | 11.00 | 7.00 | 10.00 | 9.00 | 7.00 |
| K-R20 | 0.80 | 0.78 | 0.80 | 0.84 | 0.79 | 0.77 |
| *SEM* | 2.71 | 2.73 | 2.75 | 2.72 | 2.75 | 2.76 |

[a] Combined results used for this study.

## RESULTS

The data were analyzed using the RUMM2010 program (Andrich, Sheridan, & Luo, 2000). This program was chosen for its ease of use and ability to provide item calibration and differential item functioning information in a single run.

### *Model Fit*

The first question to answer is whether or not the items are well targeted to the population being tested. With the entire placement battery taking some four hours to administer, it would *not* be an efficient use of time to require students to attempt items far above or below the majority of the students. Looking at Figure 1, it appears that the test items are bracketing the majority of the students. This is indicated by the fact that the items are distributed more or less in the same range as the students (Rasch item maps are presented in Appendixes B and C). In the future, however, it might be useful to increase the number of items of greater difficulty as the test is used to make two cutoff decisions: (a) placement into intermediate or advanced classes and (b) placement into advanced classes or exemption. As the maximum information for an item is obtained at its location estimate, increasing the number of items near the cutoff points should help give maximum information at those locations. Currently, the cutoff points are set at 50 and 60 on a scaled score (*T*-score) referenced to the norming group which corresponds roughly to difficulty estimates -0.071 and 0.619 on the scale. The approximate locations of the current cut points are indicated in Figure 1 by the dark vertical lines. Students reaching the first cut point are placed into the advanced class and students reaching the second cut point are exempted from ELI classes. It could be argued that there is no need for a good estimate of student ability below the first cut point and in a sense this is true—the intermediate class has no bottom in terms of ability. From this perspective, the three leftmost (easiest) items in Figure 1 could be eliminated without jeopardizing the decision-making information. Nevertheless, as long as the items are not degrading the instrument, there is a humane reason for including items that even the least able examinees can be successful on.

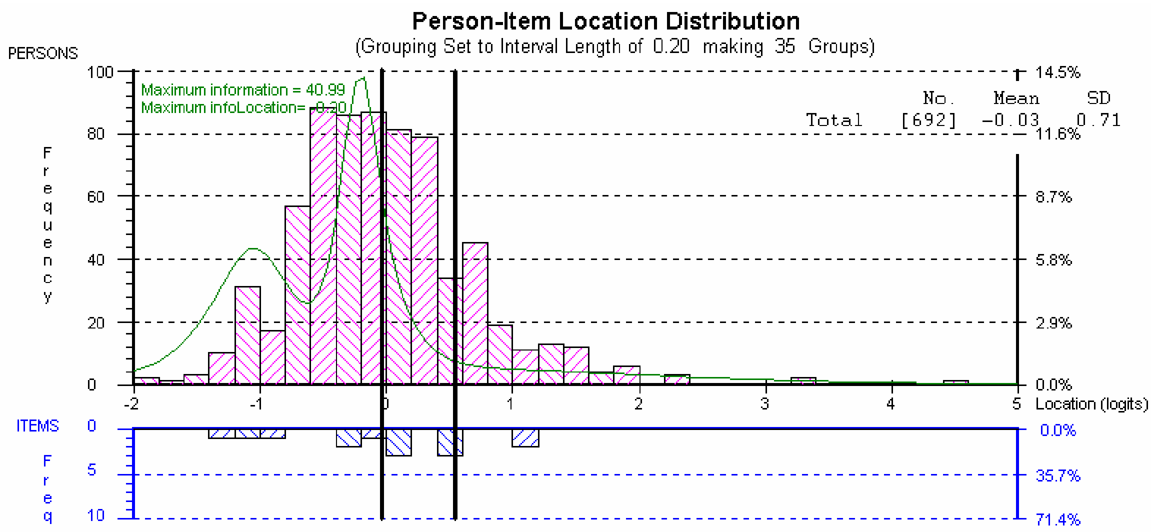

*Figure 1*. Person-Item Location Distribution with Information Function and Approximate Cut Points.

The person separation index for this test (analogous to Cronbach's alpha) was 0.76, a reasonable value and generally in line with the traditional reliability estimates produced after each administration (Table 2). Overall, the data fit the model fairly well, but there are some items that fit better than others and it is to those that we now turn.

Item fit can be assessed by comparing the actual responses of students at various ability levels to the theoretical curve predicted by the model. Table 3 shows the items on the ALT in descending order in terms of the chi-square probability. This is an approximate chi-square statistic which compares proportion of correct responses for a particular class interval with the proportion expected in the model. In this analysis, ten class intervals were used and the mean ability estimate for each class interval is compared to the theoretical proportion for that ability level. Lower probability indicates greater misfit. Though there were 692 students in the sample, one student achieved a perfect score on the ALT leaving 691 usable responses.

Table 3
*Individual Item Fit for the ALT Sorted in Order of Chi-Square Probability*

| Item Label | Location | *SE* | Residual | *df* | DatPts | ChiSq | Prob | *df* |
|---|---|---|---|---|---|---|---|---|
| D3 Voc3 | 0.030 | 0.080 | 2.653 | 639.64 | 691 | 4.507 | 0.808702 | 8 |
| P3 MLV | -0.950 | 0.040 | -0.853 | 639.64 | 691 | 4.649 | 0.794298 | 8 |
| D4 Voc4 | 0.439 | 0.082 | 2.205 | 639.64 | 691 | 7.075 | 0.528535 | 8 |
| Lec Lecture | -0.217 | 0.023 | -3.781 | 639.64 | 691 | 7.711 | 0.462181 | 8 |
| D1 Voc1 | -0.380 | 0.081 | 1.348 | 639.64 | 691 | 8.114 | 0.422400 | 8 |
| P2 Movies | -0.074 | 0.035 | 0.494 | 639.64 | 691 | 8.874 | 0.353036 | 8 |
| D5 Trans1 | 0.445 | 0.082 | 1.388 | 639.64 | 691 | 10.433 | 0.235956 | 8 |
| P1 Volcano | -1.239 | 0.046 | -0.535 | 639.64 | 691 | 12.650 | 0.124455 | 8 |
| D9 Trans5 | 1.165 | 0.092 | -0.603 | 639.64 | 691 | 14.245 | 0.075599 | 8 |
| D7 Trans3 | 0.533 | 0.083 | -0.249 | 639.64 | 691 | 15.736 | 0.046318 | 8 |
| P4 Library | -1.044 | 0.034 | -1.639 | 639.64 | 691 | 16.123 | 0.040648 | 8 |
| D2 Voc2 | 0.004 | 0.080 | -0.663 | 639.64 | 691 | 20.999 | 0.007150 | 8 |
| D8 Trans4 | 1.112 | 0.091 | 3.517 | 639.64 | 691 | 28.725 | 0.000354 | 8 |
| D6 Trans2 | 0.178 | 0.080 | 5.963 | 639.64 | 691 | 31.324 | 0.000123 | 8 |

To get a visual sense of an item showing good fit to the model, the Item Characteristic Curve (ICC) for the MLV item is presented as Figure 2. The dots on the graph represent the observed class interval averages and it is clear that they are in accord with the model predictions.

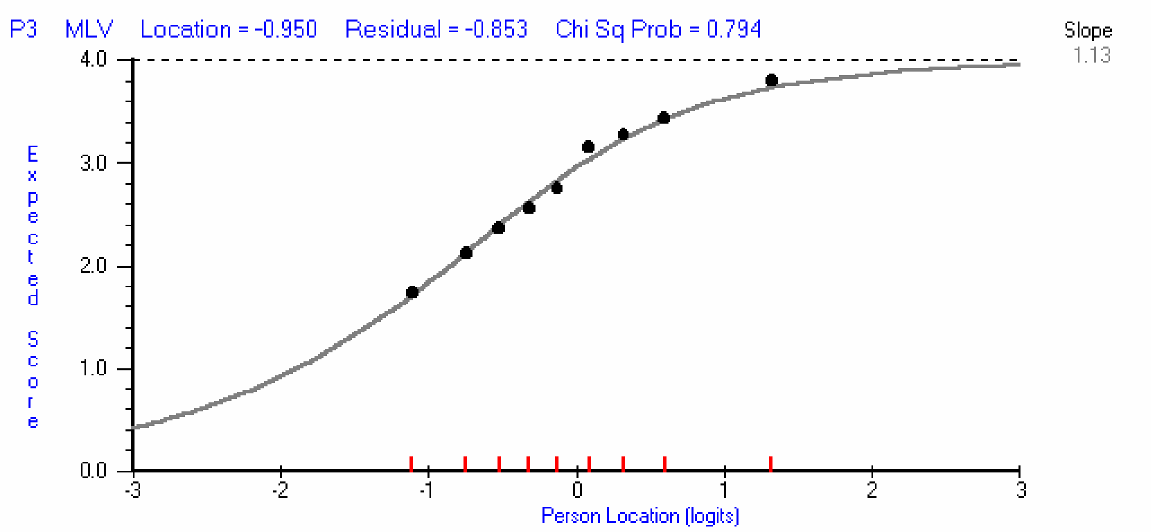

*Figure 2*. ICC for MLV Item Showing Good Model Fit.

The worst-fitting item according to the chi-square statistic was the second item in Section Three, an item that required the examinees to choose the most appropriate transition word (*whereas, not only that, for instance, likewise*) to connect two sentences that were presented aurally. The class intervals plotted on the ICC for this item, presented in Figure 3, show that success on this item does not increase consistently with an increase in ability.

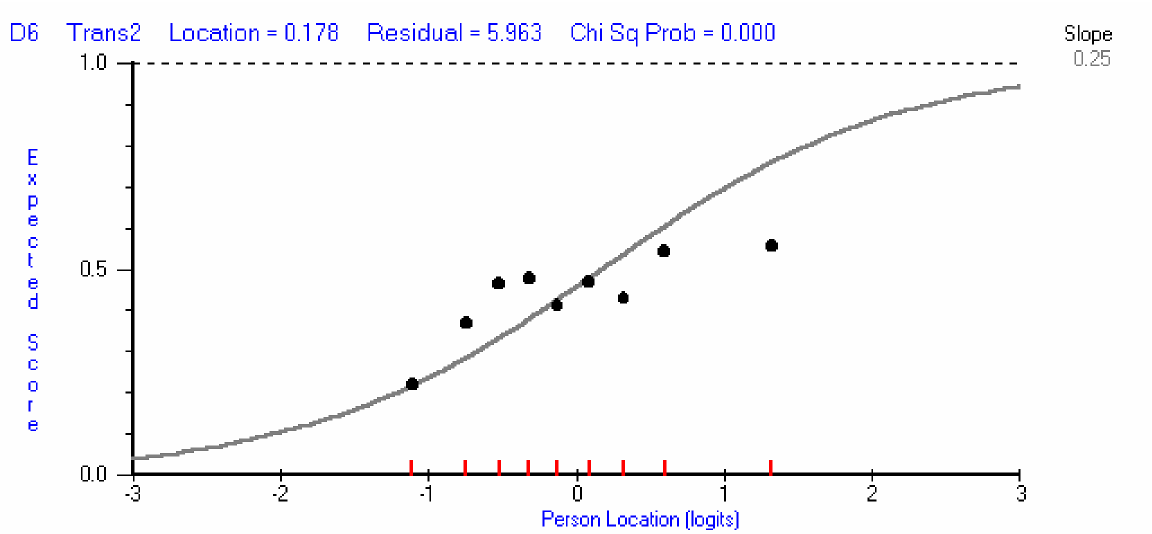

*Figure 3*. Plot of Class Intervals Against ICC for Trans 2.

For all intents and purposes, this item does not discriminate among groups of different abilities. The large positive residual value for this item (cf. Table 3) also suggests that the responses were less predictable than expected. The nature of this item was such that if one was not familiar with these words, one could probably not answer this question successfully. It is also interesting to note that the three items with the lowest chi-square probabilities were all from either Section Two or Section Three, the sections dealing with subcomponents of listening (vocabulary and transition words) rather than more global comprehension. Although these are undoubtedly an important component of listening ability, it is questionable whether this type of item is operating at the same level of scale as the listening passages. In a revised version of the test, it might be prudent to rethink the inclusion of this type of item.

It was mentioned previously that in addition to the chi-square statistic, the residual value can also be useful in interpreting possible model misfit. The Lecture passage that makes up Section Four of the ALT shows a relatively large negative residual. The class interval plot, however, shows that the observed scores are in line with their expected values (Figure 4). A negative residual value is generally interpreted as a more deterministic pattern than would be predicted by the model, in other words, the item discriminates too well. It can also be a potential sign of a local independence violation. The issue of local independence in terms of common passages has been addressed by the selection of the extended logistic model, but there might be an additional effect from the fact that students are allowed to take notes while listening to the lecture. This would make memory less of a factor for this section. This is not to suggest that the item is testing note-taking rather than listening, but rather that there could be an additional dependency in that the notes taken are also common to each response. Yen (1980) has suggested that fatigue effects can affect the discrimination of items towards the end of the test. Given that the lecture item has a 10-minute aural stimulus, this is possible. Moving the lecture item to an earlier point in the test would be one way to investigate this, but that would not be feasible as the lecture is the only passage for which students are allowed to take notes and the logistics of distributing and collecting notepaper in the middle of a test would likely create undesirable disruption.

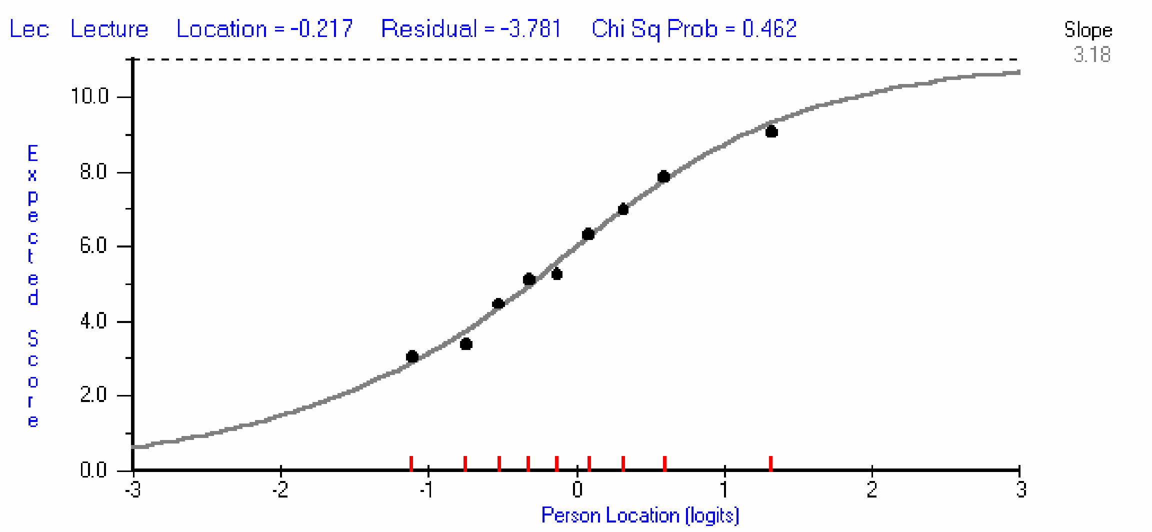

*Figure 4.* Class Interval Plots and ICC for Lecture Item.

### ***Differential Item Functioning***

Using the Rasch model, it is also possible to look at differential item functioning for different groups of students across the latent trait, and this paper will follow the approach outlined in Hagquist and Andrich (2004). By dividing the students into class intervals as before and also dividing them according to their subgroup, such as status, it is possible to generate a plot for each group separately. Essentially, the Item Characteristic Curve for any subgroup of students should not differ from the curve of the whole group. To illustrate, the ICC for the Lecture item is presented as Figure 5, with separate lines plotted for academic status. By visual inspection it is clear that the curves are essentially the same irrespective of status. This would indicate that there is no unique benefit on this item for either of the groups.

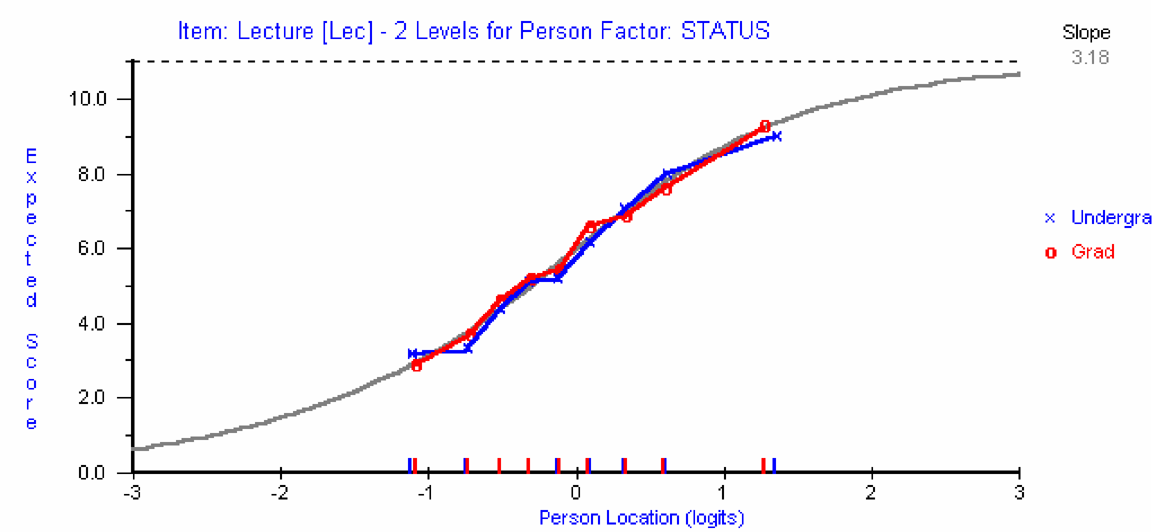

*Figure 5*. Status Plot for Lecture Showing No Differential Item Functioning.

It is possible to perform a two-way ANOVA on the residuals for class interval and status. An effect for status would indicate that item location (item difficulty) differs depending on status; an effect for class interval would indicate that the item does not fit the model across the trait irrespective of status; a significant interaction for status and class interval would indicate that the item discriminates differently for the groups. The summary of this approach for all of the items is shown in Table 4 for status and in Table 5 for language. Because of the multiple comparisons involved (14 items by 3 calculated probabilities by 2 factors), the alpha level of 0.05 will be adjusted using the Bonferroni technique to yield an alpha of 0.05 / 84 = 0.0006. Using this criterion, two items show significant DIF for status and are highlighted in Table 6. An additional item, item D8 identified previously as Trans 4, shows an effect for class interval further indicating misfit to the model.

Table 4
*Analysis of Variance of Residuals for Status with Significant Values Highlighted*

| | Status | | | Class Interval | | | Status-x-Class Interval | | |
|---|---|---|---|---|---|---|---|---|---|
| | *MS* | *F* | *p* | *MS* | *F* | *p* | *MS* | *F* | *P* |
| P1 | 5.08 | 5.784436 | 0.016434 | 1.33 | 1.515759 | 0.148044 | 0.94 | 1.074042 | 0.379325 |
| P2 | 2.97 | 3.126070 | 0.077494 | 1.12 | 1.172093 | 0.313359 | 0.63 | 0.664244 | 0.723259 |
| P3 | 0.10 | 0.111013 | 0.739095 | 0.52 | 0.581533 | 0.793541 | 0.71 | 0.802103 | 0.600873 |
| P4 | 2.00 | 2.438797 | 0.118845 | 1.93 | 2.357186 | 0.016667 | 0.42 | 0.511988 | 0.847838 |
| D1 | 2.59 | 2.673453 | 0.102511 | 1.04 | 1.073275 | 0.379870 | 0.99 | 1.021082 | 0.418423 |
| D2 | 0.69 | 0.772635 | 0.379708 | 2.67 | 3.008069 | 0.002517 | 0.78 | 0.882522 | 0.530747 |
| D3 | 0.53 | 0.527903 | 0.467741 | 0.55 | 0.549308 | 0.819385 | 2.01 | 2.002563 | 0.043797 |
| D4 | 3.79 | 3.750239 | 0.053219 | 0.89 | 0.877810 | 0.534787 | 0.41 | 0.409536 | 0.915372 |
| D5 | 0.03 | 0.033204 | 0.855440 | 1.30 | 1.360649 | 0.210568 | 2.59 | 2.709707 | 0.006094 |
| **D6** | 15.51 | 4.509630 | **0.000165** | 3.80 | 3.552106 | **0.000482** | 1.27 | 1.187211 | 0.303959 |
| D7 | 5.81 | 6.422991 | 0.011501 | 1.95 | 2.153479 | 0.029226 | 0.41 | 0.448147 | 0.891969 |
| **D8** | 0.64 | 0.582217 | 0.445709 | 4.07 | 3.728383 | **0.000290** | 3.67 | 3.367713 | 0.000851 |
| **D9** | 11.87 | 3.564560 | **0.000264** | 1.79 | 2.039708 | 0.039674 | 0.04 | 0.044811 | **N/Sig |
| Lec | 0.59 | 0.788620 | 0.374829 | 0.96 | 1.286481 | 0.247290 | 0.53 | 0.700932 | 0.690930 |

Note: See text for discussion of alpha level.

Table 5
*Analysis of Variance of Residuals for Language*

| | Language | | | Class Interval | | | Language-x-Class Interval | | |
|---|---|---|---|---|---|---|---|---|---|
| | *MS* | *F* | *p* | *MS* | *F* | *p* | *MS* | *F* | *p* |
| P1 | 0.48 | 0.544707 | 0.651860 | 1.33 | 1.509628 | 0.150236 | 1.03 | 1.165110 | 0.266891 |
| P2 | 1.50 | 1.586126 | 0.191461 | 1.12 | 1.181321 | 0.307649 | 1.06 | 1.126942 | 0.307016 |
| **P3** | 5.64 | 6.706263 | **0.000181** | 0.52 | 0.613116 | 0.767260 | 1.45 | 1.721433 | 0.017820 |
| P4 | 3.67 | 4.588174 | 0.003443 | 1.93 | 2.417798 | 0.014090 | 0.94 | 1.176825 | 0.255300 |
| D1 | 0.95 | 0.974034 | 0.404463 | 1.04 | 1.073036 | 0.380086 | 1.04 | 1.073668 | 0.368780 |
| D2 | 0.97 | 1.091642 | 0.351945 | 2.67 | 2.996203 | 0.002618 | 0.74 | 0.828008 | 0.702244 |
| **D3** | 7.32 | 7.374140 | **0.000088** | 0.55 | 0.554012 | 0.815667 | 0.76 | 0.763775 | 0.784158 |
| D4 | 0.71 | 0.698661 | 0.553085 | 0.89 | 0.876013 | 0.536338 | 0.91 | 0.897454 | 0.606334 |
| **D5** | 5.83 | 6.156807 | **0.000396** | 1.30 | 1.372911 | 0.205013 | 1.08 | 1.145862 | 0.286676 |
| D6 | 4.40 | 4.115318 | 0.006607 | 3.80 | 3.551384 | 0.000475 | 1.31 | 1.229985 | 0.207045 |
| D7 | 0.41 | 0.452251 | 0.715791 | 1.95 | 2.142274 | 0.030151 | 0.87 | 0.962405 | 0.515159 |
| D8 | 2.82 | 2.571817 | 0.053197 | 4.07 | 3.716109 | **0.000278** | 1.62 | 1.479276 | 0.066211 |
| D9 | 2.81 | 3.240184 | 0.021710 | 1.79 | 2.059716 | 0.037662 | 1.02 | 1.175691 | 0.256404 |
| **Lec** | 4.74 | 6.573551 | **0.000217** | 0.96 | 1.335624 | 0.222503 | 0.92 | 1.276178 | 0.170816 |

Note: See text for discussion of alpha level.

Although the Lecture item showed no DIF for status, it does show potential DIF for language, as can be clearly seen when the curves for each language are plotted separately in Figure 6. Though language effects for vocabulary or structure can be potentially attributed to the presence or absence of L1 cognates among other things, there is no immediate explanation for why language background should be relevant for this item. As this item represents possibly the most authentic criterion task – listening to an extended piece of discourse for general understanding – it is prudent to subject it to further examination rather than to slate it for deletion.

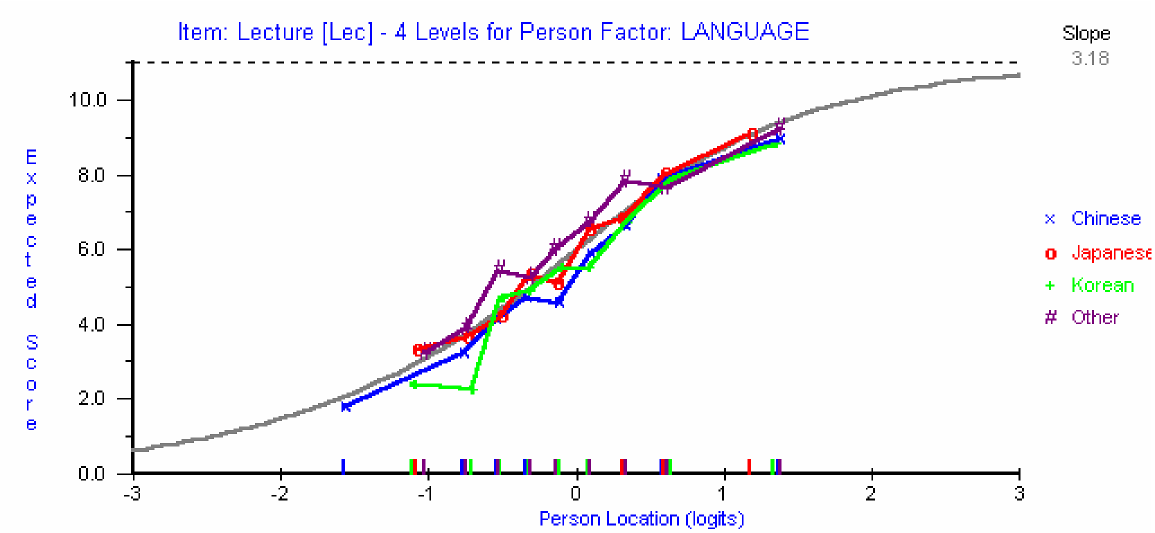

*Figure 6.* Lecture Item Showing Differential Item Functioning.

### *Individual Person Fit*

The majority of this paper has been spent discussing item fit and functioning on the ALT and no mention has been made of the fit of persons to the model. This may seem like a grievous oversight as students, not items, are the ones being placed. The fact is, however, that the better the items fit the model, the more useful student misfit information is as the properties of the items have been confirmed.

Before looking at particular individual student performance, it is important to consider how test scores are used in placement decisions. Though I have been referring to the ALT as *the* placement test for the purposes of this paper, in actuality three scores are generally considered in placing the students into ELI listening classes– the ALT score, the score from a dictation test given as part of the ELI test battery, and the TOEFL section score for listening. The two ELI test scores are considered first, and the TOEFL score is used if needed. The philosophy at the ELI is to give the student the benefit of the doubt, and if the different test scores would indicate different placements, the highest of the three scores is used to make the placement decision. In addition, the standard error of measurement is considered if two or more scores are near the cut point.

Even with this relatively cautious approach, however, the underlying assumption that the total score on the test is a valid indicator of the student's ability is not questioned.

Under the Rasch paradigm, the total score is only a useful measure of underlying ability if the item responses of the student fit the model. Therefore, students whose response patterns don't fit the model should be flagged for potential follow-up or alternative assessment procedures. This approach has been proposed for college admissions (Tognolini & Andrich, 1996) in which many different indicators from each student's profile are aggregated to form a single score which is then used for final decisions. Since the sheer number of applicants prohibits an in-depth consideration of each student, individual indicators can be treated as analogous to test items and those students whose profiles do not fit the model can be examined in more depth. Though this is not presently done in the ELI, it would be possible using a Rasch approach to placement.

Misfit for individuals is usually indicated by a large negative or positive standardized residual value. Using the generally accepted value of ±2.00, only a few students showed misfit to the model. To give a sense of the information available under the Rasch model, one student is highlighted here. Summary results are presented in Appendix D. Table 6 shows the student's responses to each item. The student's ability estimate was 0.371. Therefore, one would expect that she should be successful on items below that level. However, this wasn't the case for the first passage, Item P1. Though the item was a good distance below her ability level, this student only achieved a score of 1 when the model expected score was 3.5. Perhaps this person was a little slow in getting started or a little nervous when listening to the first passage.

Table 6

Person by item fit for Student 498

```
Serial Number: 498
ID             Stu498
Gender         Female     (Code:   2)
Language       Japanese   (Code:   2)
Status         Grad       (Code:   2)
-----------------------------------------------------------------------
 Item        Item                    Item    Obs     Exp          Std
 Code        Statement               Locn   Score   Score        Resid
-----------------------------------------------------------------------
 P1    Volcano                     -1.239     1     3.508       -3.756  #
 P2    Movies                      -0.074     3     3.465       -0.403
 P3    MLV                         -0.950     2     3.269       -1.476
 P4    Library                     -1.044     6     5.155        0.918
 D1    Voc1                        -0.380     0     0.680       -1.456
 D2    Voc2                         0.004     1     0.591        0.832
 D3    Voc3                         0.030     1     0.585        0.843
 D4    Voc4                         0.439     0     0.483       -0.966
 D5    Trans1                       0.445     1     0.482        1.038
 D6    Trans2                       0.178     1     0.548        0.908
 D7    Trans3                       0.533     0     0.460       -0.922
 D8    Trans4                       1.112     0     0.323       -0.690
 D9    Trans5                       1.165     0     0.311       -0.672
 Lec   Lecture                     -0.217    11     7.142        2.236  *
-----------------------------------------------------------------------
Person:    Location =   0.371     SE =  0.344     Fit        =   2.913
Sample:    Mean     =  -0.030     SE =  0.710     Sep. Index =   0.760
-----------------------------------------------------------------------
```

Because the listening passage items were composed of several individual questions from which an overall item score was achieved, it is also instructive to look at the same student's responses in terms of the threshold for each item. This is shown in Figure 7. The student's ability estimate is shown as a dark vertical line and the narrower lines on either side indicate the confidence interval for that estimate. Here it is clear that a score of 1 on the Volcano passage and a score of 11 on the lecture are clearly seen as anomalous for a student of her ability. It is important to keep in mind that the thresholds represented here are the thresholds between different scores and do not correspond to success on particular items for the passages. Different students may achieve the same threshold through different combinations of items.

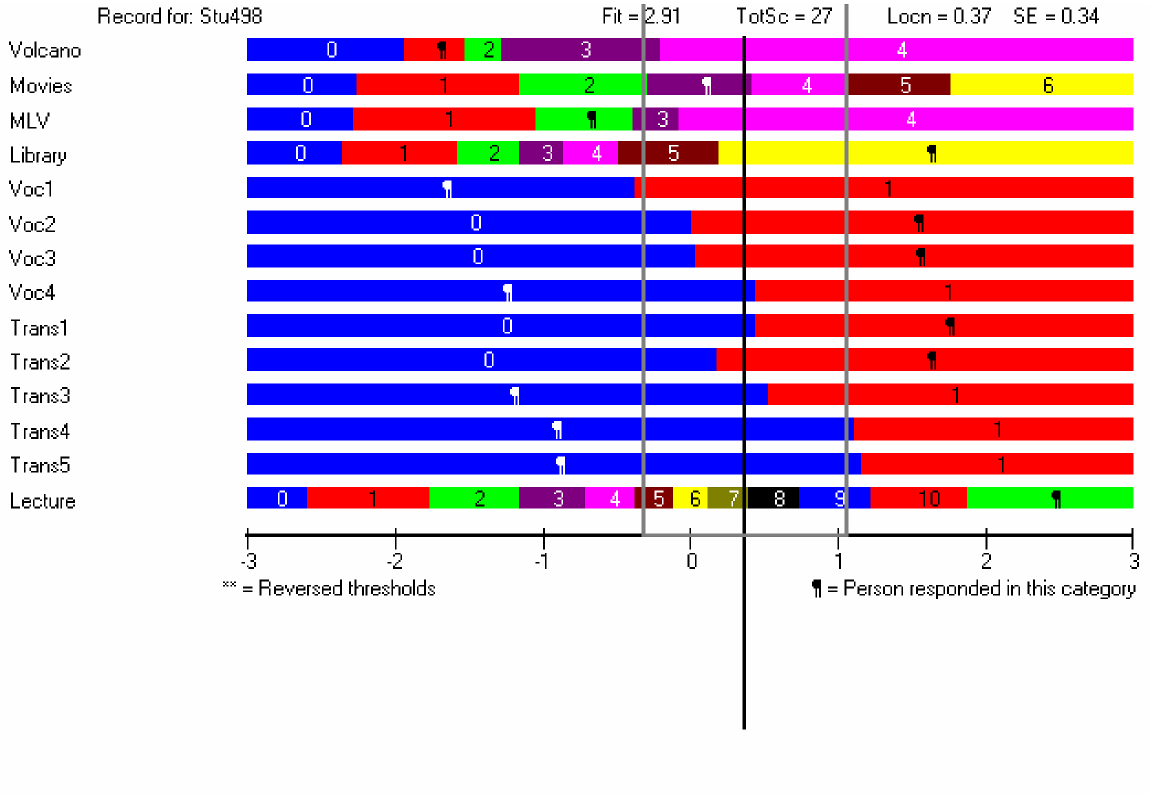

*Figure 7*. Threshold Map and Response Pattern for Student 498.

## CONCLUSIONS

This paper presented an analysis of the Academic Listening Test (ALT) used by the University of Hawaii at Manoa of Hawaii at Manoa for placing incoming students into listening classes. The desirable qualities for a good placement instrument were discussed and it was argued that a test fitting the Rasch model would have these qualities. The ALT was examined from this perspective and it was found that, although the test shows reasonable fit to the model, there is room for improvement. The items generally bracketed the range of student ability, but the test information was not maximized at the decision points. It was also noted that some of the items showed some misfit to the model, especially those that dealt with subcomponents of listening. It was suggested earlier that perhaps the reason for this misfit is that items in these sections are not at the same level of scale as rest of the test. The ALT was designed as a test of academic listening ability and the short passages and lecture seem to be measuring that ability at a more global level than the problematic sections. It was suggested that in a revision of the test it would perhaps be prudent to replace these problematic sections with additional passages also geared towards global comprehension. In addition, differential item functioning was found for some of the items for language background or status, and those items deserve closer scrutiny. Finally, the advantage of using Rasch misfit information to consider the performance of individual students was presented. Assuming that the test is machine-scored, the availability of fairly user-friendly software does realistically allow for the generation and use of student misfit statistics even within the most restrictive time constraints.

## REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. B. Tuma (Ed.), *Social Methodology* (pp. 33-80). San Fransisco: Jossey-Bass.

Andrich, D. (1988a). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education, 1*(4), 363-378.

Andrich, D. (1988b). *Rasch models for measurement*. Newberry Park, CA: SAGE Publications, Inc.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath & S. H. Lovibond (Eds.), *Proceedings of the XXIVth International Congress of Psychology. Mathematical and theoretical systems* (Vol. 4, pp. 7-16). North Holland: Elsevier Science Publications.

Andrich, D. (2002a). A framework relating outcomes based education and the taxonomy of educational objectives. *Studies in Educational Evaluation, 28*, 35-59.

Andrich, D. (2002b). Implications and applications of modern test theory in the context of outcomes based education. *Studies in Educational Evaluation, 28*, 103-121.

Andrich, D., Sheridan, B., & Luo, G. (2000). RUMM2010: A Windows interactive program for analyzing data with Rasch Unidimensional Models for Measurement [computer program]. Perth, Western Australia: RUMM Laboratory.

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

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

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

Blais, J.-G., & Laurier, M. D. (1995). The dimensionality of a placement test from several different analytical perspectives. *Language testing, 12*(1), 72-98.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Canale, M. (1983). From communicative competence to communicative language pedagogy. In R. W. Schmidt (Ed.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: a tentative model for test specification and development. *The Modern Language Journal, 77*(2), 180-191.

Hagquist, C., & Andrich, D. (2004). Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences, 36*(4), 955-968.

Kondo-Brown, K., & Brown, J. D. (2000). *The Japanese placement tests at the University of Hawai'i: Applying item response theory* (NFLRC Net Work #20). Honolulu: University of Hawai'i, Second Language Teaching & Curriculum Center. http://www.nflrc.hawaii.edu/NewWorks/NW18/[Retrieved September 16, 2003].

Linacre, J. M. (1992). Why fuss about statistical sufficiency? *Rasch Measurement Transactions, 6*(3), 230. Available: http://209.238.226.290/rmt/rmt263c.htm.

Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education, 1*(4), 279-297.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. New York: Cambridge University Press.

Rasch, G. (1960/80). *Probabalistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: The University of Chicago Press.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placment test. *Language testing, 8*(2), 95-111.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Tognolini, J., & Andrich, D. (1996). Analysis of profiles of students applying for entrance to universities. *Applied Measurement in Education, 9*(4), 323-353.

van der Linden, W. J. (1992). Sufficient and necessary statistics. *Rasch Measurement Transactions, 6*(3), 231. Available: http://209.238.226.290/rmt/rmt263d.htm.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues & Practice, 16*(4), 33-45.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*(4), 297-311.

## APPENDIX A

## FREQUENCY OF SELF-REPORTED FIRST LANGUAGE

| Language | $N$ |
|---|---|
| Assamese | 1 |
| Bengali | 2 |
| Bosnian | 1 |
| Cantonese | 24 |
| Chamorro | 1 |
| Chinese | 101 |
| Chuukee | 2 |
| English | 3 |
| Filipino | 2 |
| Finnish | 1 |
| French | 13 |
| German | 7 |
| Greek | 1 |
| Hebrew | 2 |
| Hungarian | 1 |
| Ilokano | 4 |
| Indonesian | 10 |
| Japanese | 283 |
| Khmer | 1 |
| Korean | 120 |
| Kosraean | 1 |
| Laotian | 2 |
| Malay | 2 |
| Mandarin | 12 |
| Marhallese | 1 |
| Micronesian | 2 |
| Mongolian | 2 |
| Nepali | 1 |
| Oriya | 1 |
| Palauan | 1 |
| Pohnpeian | 3 |
| Polish | 2 |
| Portuguese | 5 |
| Romanian | 1 |
| Russian | 3 |
| Samoan | 9 |
| Serbian | 1 |
| Spanish | 7 |
| Swedish | 2 |
| Tagalog | 2 |
| Tamic | 1 |
| Telugu | 1 |
| Tetun | 7 |

| | |
|---|---|
| Thai | 27 |
| Tongan | 1 |
| Vietnamese | 15 |
| TOTAL | 692 |

# APPENDIX B

## RASCH ITEM MAP SHOWING PERSON-ITEM LOCATION ESTIMATES

```
------------------------------------------------------------------------------
LOCATION          PERSONS     ITEMS [locations]
------------------------------------------------------------------------------
  5.0                      |
                           |
                           |
                           |
                           |
  4.0                      |
                           |
                           |
                           |
                           |
  3.0                      |
                           |
                           |
                           |
                         X |
  2.0                      |
                         X |
                         X |
                        XX |
                       XXX |
  1.0                   XX |    D8     D9
                      XXXX |
                 XXXXXXXXX |
                   XXXXXXX |    D4     D5     D7
          XXXXXXXXXXXXXXXX |
  0.0     XXXXXXXXXXXXXXXX |    D2     D3     D6
         XXXXXXXXXXXXXXXXX |    P2
         XXXXXXXXXXXXXXXXX |    D1    Lec
        XXXXXXXXXXXXXXXXXX |
               XXXXXXXXXXX |
 -1.0                  XXX |    P3
                    XXXXXX |    P4
                        XX |    P1
                         X |
                           |
 -2.0                      |
------------------------------------------------------------------------------
            X = 5 Persons
------------------------------------------------------------------------------
```

# APPENDIX C

## RASCH ITEM MAP SHOWING PERSON-THRESHOLD ESTIMATES

```
------------------------------------------------------------------------------
LOCATION          PERSONS     ITEMS [uncentralised thresholds]
------------------------------------------------------------------------------
  5.0                      |
                           |
                           |
                           |
                           |
  4.0                      |
                           |
                           |
                           |
                           |
  3.0                      |
                           |
                           |
                           |
                         X |
  2.0                      |
                         X |   Lec.11
                         X |    P2.06
                        XX |
                       XXX |   Lec.10
  1.0                   XX |    P2.05     D8.01     D9.01
                      XXXX |
                 XXXXXXXXX |   Lec.09
                   XXXXXXX |    P2.04     D4.01     D5.01     D7.01
          XXXXXXXXXXXXXXXX |   Lec.08
  0.0     XXXXXXXXXXXXXXXX |    D2.01     D3.01    Lec.07     D6.01     P4.06
         XXXXXXXXXXXXXXXXX |   Lec.06     P3.04
         XXXXXXXXXXXXXXXXX |    P3.03     D1.01    Lec.05     P2.03     P1.04
        XXXXXXXXXXXXXXXXXX |    P4.05
               XXXXXXXXXXX |   Lec.04
 -1.0                  XXX |    P4.04
                    XXXXXX |    P4.03    Lec.03     P2.02     P3.02
                        XX |    P1.03
                         X |    P4.02     P1.02
                           |   Lec.02
 -2.0                      |    P1.01
                           |
                           |    P4.01     P3.01     P2.01
                           |
                           |   Lec.01
 -3.0                      |
------------------------------------------------------------------------------
            X = 5 Persons
------------------------------------------------------------------------------
```

## APPENDIX C

### PERSON FREQUENCY DISTRIBUTION SORTED BY RESIDUAL

| GROUP | RESIDUAL | FREQUENCY | CumFREQ | CumFREQ% |
|---|---|---|---|---|
| 1 | < -3.00 | 0 | 0 | 0.00 |
| 2 | -3.00 to -2.80 | 0 | 0 | 0.00 |
| 3 | -2.80 to -2.60 | 0 | 0 | 0.00 |
| 4 | -2.60 to -2.40 | 1 | 1 | 0.14 |
| 5 | -2.40 to -2.20 | 3 | 4 | 0.58 |
| 6 | -2.20 to -2.00 | 3 | 7 | 1.01 |
| 7 | -2.00 to -1.80 | 2 | 9 | 1.30 |
| 8 | -1.80 to -1.60 | 7 | 16 | 2.31 |
| 9 | -1.60 to -1.40 | 22 | 38 | 5.49 |
| 10 | -1.40 to -1.20 | 27 | 65 | 9.39 |
| 11 | -1.20 to -1.00 | 27 | 92 | 13.29 |
| 12 | -1.00 to -0.80 | 47 | 139 | 20.09 |
| 13 | -0.80 to -0.60 | 51 | 190 | 27.46 |
| 14 | -0.60 to -0.40 | 50 | 240 | 34.68 |
| 15 | -0.40 to -0.20 | 67 | 307 | 44.36 |
| 16 | -0.20 to 0.00 | 61 | 368 | 53.18 |
| 17 | 0.00 to 0.20 | 60 | 428 | 61.85 |
| 18 | 0.20 to 0.40 | 53 | 481 | 69.51 |
| 19 | 0.40 to 0.60 | 55 | 536 | 77.46 |
| 20 | 0.60 to 0.80 | 47 | 583 | 84.25 |
| 21 | 0.80 to 1.00 | 39 | 622 | 89.88 |
| 22 | 1.00 to 1.20 | 27 | 649 | 93.79 |
| 23 | 1.20 to 1.40 | 17 | 666 | 96.24 |
| 24 | 1.40 to 1.60 | 13 | 679 | 98.12 |
| 25 | 1.60 to 1.80 | 5 | 684 | 98.84 |
| 26 | 1.80 to 2.00 | 4 | 688 | 99.42 |
| 27 | 2.00 to 2.20 | 1 | 689 | 99.57 |
| 28 | 2.20 to 2.40 | 1 | 690 | 99.71 |
| 29 | 2.40 to 2.60 | 0 | 690 | 99.71 |
| 30 | 2.60 to 2.80 | 0 | 690 | 99.71 |
| 31 | 2.80 to 3.00 | 1 | 691 | 99.86 |
| 32 | > 3.00 | 1 | 692 | 100.00 |