Evaluating and customizing a commercially produced EFL placement examination

Christopher Weaver
Toyo University

Andrew Jones
Juergen Bulach
Jissen Women's University

Reference data:

Weaver, C., Jones, A., & Bulach, J. (2008). Evaluating and customizing a commercially produced EFL placement examination. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*. Tokyo: JALT.

This paper reports on the use of Rasch measurement theory to evaluate and customize a commercially produced EFL multiple-choice placement examination. The evaluation process revealed which types of items provided the most information about students' current level of English proficiency. These findings in turn led to suggestions on how to customize the placement examination to meet the specific assessment needs of a university.

本論文では、市販のEFLプレースメントテストを評価し、カスタマイズすることを目的とした、ラッシュ測定理論の利用に関する報告を行なう。評価の過程を観察することにより、受験生の英語能力レベルについて、どのような質問項目が最も有用な情報を提供し得るのか、ということを明らかにした。この結果は、大学のアセスメントニーズに合致したプレースメントテストのカスタマイズを考える際の、非常に大きな示唆となるものであった。

A number of universities in Japan have begun to use placement examinations to stream new students into their language programs. In principle, placement examinations can be a very effective means of ensuring that students receive instruction suited to their current level of language competence.

The benefits of streaming students, however, rest upon the quality of the placement examination. Rasch measurement theory (Rasch, 1960/1980) provides test writers and administrators responsible for placement decisions with a number of analytical strategies that can help ensure the highest possible level of measurement accuracy.

In the context of the current investigation of a multiple-choice EFL placement examination, the Rasch dichotomous rating scale model transforms the students' responses on the placement examination into two types of estimates. The first estimate is the students' level of English proficiency. The second estimate is the level of difficulty that each item has on the placement examination. These estimates are reported in a unit of measurement called the logit, which is best thought of as a measure of the probability of a student correctly answering the different items on the placement examination. These estimates also serve as a basis for evaluating the performance of the placement examination. Unlike other types of statistical models such as item response theory, which attempts to accurately model students' responses, a Rasch analysis starts with specific measurement properties that are used to determine the extent to which the placement examination exhibits sound measurement. In other words, the Rasch model is not seeking to accurately model data, but rather the objective of a Rasch analysis is to seek data that fits the model (Andrich, 1989). This objective thus gives rise to a number of different analytical strategies which can be used to evaluate and customize a commercially produced EFL placement examination.

The first Rasch-based strategy that test writers and administrators can use to evaluate the performance of a placement examination involves monitoring the amount of overlap that exists between the range of students' level of English proficiency and the difficulty range of the items on the placement examination. If there is a significant amount of overlap between the two, the placement examination does a good job at *targeting* the student population. Targeting is important because items that have a level of difficulty close to students' level of proficiency provide the greatest amount of statistical information about students while reducing the amount of measurement error (Wright & Stone, 1979). A concentration of items located around the probable cut point(s) where students are streamed into classes of different levels is equally important in order to ensure the most accurate and reliable information possible for placement decisions. This type of target is known as *cut-point targeting* (Weaver & Sato, in press).

The second strategy involves a distractor analysis that examines the average proficiency estimates of the students who choose the different options of a multiple-choice item (Bond & Fox, 2007). From the Rasch perspective, a multiple-choice item is performing well when students who have chosen the correct option have a higher average proficiency estimate than the students who chose one of the incorrect options. In terms of the incorrect options, they are working well if they attract students who have different levels of English proficiency. Thus, a well performing multiple-choice item has options (i.e. the correct response and the distractors) that can discern a wide range of student proficiencies.

The third strategy involves examining the model fit for the different items on the placement examination. Model fit is the degree of agreement between the Rasch model's expectations of how an item should perform on the placement examination and how the item actually performs. For example, the Rasch model expects that when a student's level of English proficiency is equal to an item's level of difficulty, the student has a fifty percent chance of correctly answering the item. When this expectation is not met, the result is a large fit statistic. Typically, outfit and infit statistics are used to evaluate model fit. The outfit statistic is an outlier-fit statistic sensitive to unexpected behavior far away from the person's level of proficiency (Linacre, 2004). For example if a group of low level proficiency students has successfully answered a very difficult item on the placement examination, the item will have a large outfit statistic because the Rasch model expects that this group of low level students should have answered the item incorrectly. The infit statistic, in contrast, is an information-weighed fit statistic sensitive to unexpected behavior close to the person's level of proficiency. In other words, if a group of students whose level of proficiency is close to an item's level of difficulty unexpectedly answer the item incorrectly, the item will have a large infit statistic because the Rasch model expects these students to have a fifty percent chance of correctly answering the item.

The frame of reference for the outfit and infit statistics for this investigation was determined with simulated data that fit the Rasch model. This simulated data was based on the distribution of item and person estimates from a calibration of the real data. The standard deviation for the infit and outfit statistics was 0.7 based upon the simulated data set, which was then multiplied by two to provide a benchmark yielding an approximate Type I error rate of 5%. Thus, items with outfit and/or infit statistics exceeding ±1.4 were considered to be contributing more off-variable noise than useful information.

Arising from these three Rasch-based strategies are the following research questions that guided this investigation of a commercially produced EFL placement examination:

1. To what extent do the items on the placement examination target the students' English proficiency?
2. To what extent do the different sections of the placement examination define a meaningful continuum of student proficiency?
3. How can items on the placement examination be customized to improve the performance of the examination and, more importantly, to help improve placement decisions?

Method

Participants

The evaluation of the placement examination involved the tests of 2,161 female students attending a private women's university located on the outskirts of Tokyo, Japan over a three-year period. The longitudinal nature of this data helps ensure that the results of the evaluation were more representative of students' English proficiency past and present. The students included English and non-English majors who are required to take a semester-long English communication course. The course meets twice a week with one meeting being taught by a group of native speakers of

English using the Student's Book section of *Full Contact 1A* (Richards, Hull, Proctor, & Shields, 2006) and the second meeting being taught by a group of Japanese teachers of English using the Video Activity Book section of *Full Contact 1A*.

The placement examination

The placement examination is a commercially produced test originating from the Placement and Evaluation Package (Lesley, Hansen, & Zukowski-Faust, 2003, p. 49-66) prepared for *New Interchange* series. The examination has three sections. The first section focuses on listening. This section has 20 items that initially correspond to different spoken conversations involving two speakers. Yet as the students progress through the listening section, a greater number of items correspond to one spoken conversation. The length of the spoken interactions also increases in duration. Students have 15 minutes to complete the listening section of the placement examination.

The second section of the placement examination focuses upon students' reading skills. This section has 20 items. Similar to the listening section, the number of items corresponding to each reading passage and the length of the reading passages increase as the students progress through the reading section of the placement examination. Students have 20 minutes to complete this section of the examination.

The third section of the placement examination focuses upon language use. This section has 30 discrete items designed to assess students' grammatical competence. Students have 15 minutes to complete this section.

All of the items on the placement examination are multiple-choice with four possible responses. Students write their answers on a mark sheet, which is machine-scored immediately after the placement examination.

Procedure

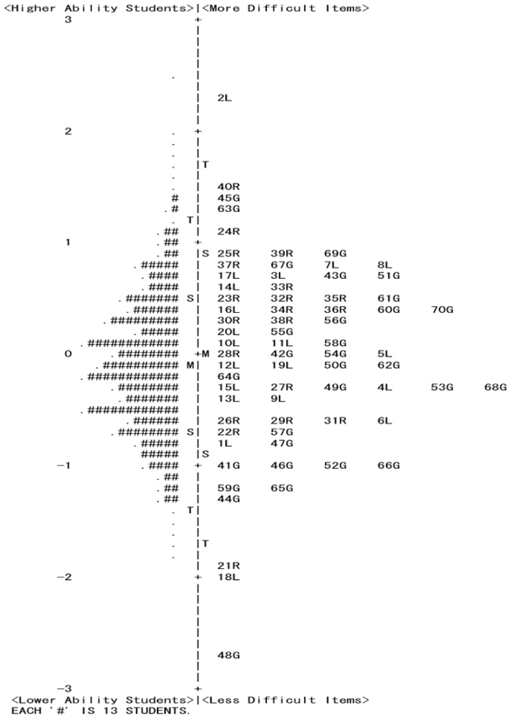
On the first day of the classes, the students take the placement examination. The results of the examination are then used to identify the highest proficiency students in order to create a challenge class for each department. The remaining students are then assigned in alphabetical order to classes for their department.

Results

Performance of the entire placement examination

The performance of the entire placement examination is graphically illustrated using a Wright map (Figure 1) produced by Winsteps (Linacre, 2007). This graphical output is basically two standard distribution curves turned vertically and then brought together. The left side of the Wright map is the standard distribution curve of the students' level of English proficiency. In other words, it is the ranking of the 2,161 Japanese students based upon their responses to the 70 items on the placement examination. Students with a higher level of English proficiency are located on the upper left-hand side of the Wright map. Students who have a lower level of English proficiency are located on the lower left-hand side. When interpreting the students' locations on the Wright map, it should be remembered that, in the case of this analysis, each number sign (#) represents thirteen students and a dot (.) represents less than thirteen students. In addition, the "M" marker on the left side of the map indicates the mean proficiency level or the average level of English proficiency for these students. The "S" and the "T" are place markers for standard deviation. The "S" markers specify one standard deviation above and below the mean. The "T" markers are placed two standard deviations away from the mean.

The right side of the Wright map is the standard distribution curve for the 70 items on the placement examination based upon their level of difficulty. The item difficulty reflects the level of difficulty which the 2,161 Japanese students had in choosing the correct response for each item. More difficult items on the placement examination are located on the upper right-hand side of the Wright map, whereas less difficult items are located on the lower right-hand side. Each item on the placement examination also has a number and letter code to assist in the interpretation of the results. The number indicates the sequence of the items on the placement examination. The letter indicates to which section of the placement examination the item belongs (e.g. listening, reading, or language use). For example, item 2L is the second item on the placement examination and it appears in the listening section; item 40R is the fortieth item and it is from the reading section; and item 45G is the forty-fifth item of the placement examination and it belongs to the language use (grammar) section. These three items are the most difficult items on the placement examination and thus they are located on the upper-right hand side of the Wright map.



```
<Higher Ability Students>|<More Difficult Items>
    3                    +
                         |
                         |
                         |
                         |
                      .  |
                         |
                         |   2L
                         |
                         |
    2                 .  +
                      .  |
                      .  |
                      .  |T
                      .  |
                      .  |   40R
                      #  |   45G
                    . #  |   63G
                      . T|
                   . ##  |   24R
    1              . ##  +
                   . ##  |S  25R    39R    69G
                . #####  |   37R    67G    7L     8L
                 . ####  |   17L    3L     43G    51G
                 . ####  |   14L    33R
              . ####### S|   23R    32R    35R    61G
               . ######  |   16L    34R    36R    60G    70G
            . #########  |   30R    38R    56G
                . #####  |   20L    55G
        . #############  |   10L    11L    58G
    0        . ########  +M  28R    42G    54G    5L
           . ########## M|   12L    19L    50G    62G
        . #############  |   64G
             . ########  |   15L    27R    49G    4L     53G    68G
              . #######  |   13L    9L
        . #############  |
               . ######  |   26R    29R    31R    6L
             . ######## S|   22R    57G
                . #####  |   1L     47G
                  #####  |S
   -1            . ####  +   41G    46G    52G    66G
                   . ##  |
                   . ##  |   59G    65G
                   . ##  |   44G
                      . T|
                      .  |
                      .  |
                      .  |T
                      .  |
                         |   21R
   -2                    +   18L
                         |
                         |
                         |
                         |
                         |
                         |
                         |   48G
                         |
                         |
   -3                    +
 <Lower Ability Students>|<Less Difficult Items>
 EACH '#' IS 13 STUDENTS.
```

Figure 1. Wright map of the entire placement examination

The "M" marker on the right side of the map indicates the average level of difficulty for the 70 items on the placement examination. Once again, the "S" and "T" markers specify the standard deviations above and below the mean. The location of the two "M" markers on the Wright map indicates that the average level of item difficulty of the placement examination is slightly higher than the average proficiency of the students over the three-year period. Yet, overall, the difficulty level of the examination targets the students' level of English proficiency. The dash-line box on Figure 1 shows that item 2L is the only item on the placement examination which is beyond the students' level of English proficiency. On the opposite end of the continuum, there are three items (i.e. 21R, 18L, and 48G) that have a level of difficulty below the students' level of English proficiency.

Performance of the different sections of the placement examination

The performance of the different sections of the placement examination can be evaluated using a slightly modified Wright map (Figure 2). This graphical output groups the items from the different sections of the placement examination into three separate columns on the right side of the Wright map.

The first column shows the listening section items on the placement examination. The difficulty level of these 20 items spans 4.36 logits starting with the easiest item 18L located at -2.03 logits and ending with the most difficult item 2L located at 2.33 logits. There is a considerable space between these two items and the rest of the questions on the listening section of the examination. This spacing suggests that there are some gaps where there are not enough items to accurately define students' English listening proficiency at the lower and the higher ends of the continuum.

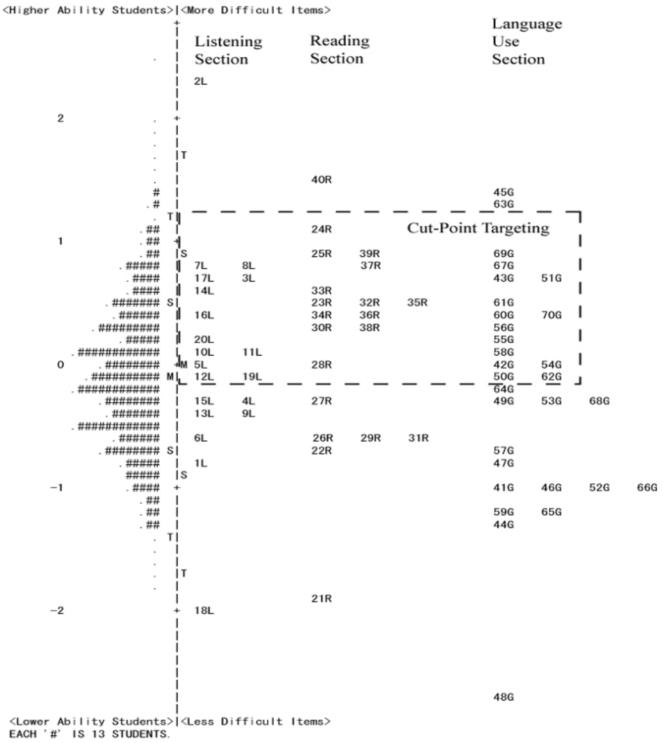


Figure 2. Wright map of the different sections of the placement examination

Interestingly, the difficulty level of the items on the listening section is not sequential. In other words, increasing the number of items that students must answer and increasing the length of the spoken dialogues do not necessarily result in an increasing level of item difficulty.

As a result, other factors such as the cognitive processing required of the different listening items might provide a possible explanation for the non sequential order.

The second column shows the reading section items on the examination. The difficulty level of these 20 items spans 3.34 logits starting with the easiest item 21R located at -1.88 logits and ending with the most difficult item 40R located at 1.46 logits. Along this continuum of item difficulty, there are a number of locations where there are either no or few items defining students' level of reading proficiency. The lack of items is especially apparent in the proficiency range from -0.64 to 0.28 logits. There are only two items (i.e. 27R and 28R) that can accurately define the level of English proficiency of students who are located around the mean.

The sequence of the reading section items generally follows a pattern in which the items progressively become more difficult. Increasing the number of items that students need to answer and increasing the length of the reading passages that students need to read seem to be two factors that mediate these reading section items' level of difficulty.

The third column shows the language use section items on the placement examination. The difficulty level of these 30 items spans 4.12 logits starting with the easiest item 48G located at -2.69 logits and ending with the most difficult item 45G located at 1.43 logits. There are a couple of gaps between items along the continuum of students' level of English proficiency. A notable gap is between items 49G and 57G where there are over 325 students who are undifferentiated by the items in the language use section. Another apparent gap is at the lower end of the grammar proficiency range between items 48G and 44G.

Gaps along the continuum of students' level of English proficiency, however, must be considered in relation to the purpose of the placement examination. This examination aims to identify high proficiency students in order to create a challenge class for each department. As such, cut-point targeting becomes important because there is a need to have enough items on the placement examination that can accurately detect differences amongst students who are located around the probable cut-point. The dashed-line box in Figure 2 shows that 39 items (i.e. 12 listening items, 13 reading items, and 14 language use items) are located between the "M" marker (i.e. the mean proficiency level for the students) and the "T" marker (i.e. two standard deviations above the mean) for the students' level of English proficiency over a three-year period. Having enough items to clearly define this relatively large range of student proficiency is necessary because students from the different departments at the university have considerably different levels of English proficiency (Weaver, 2008). As a result, the cut point for each department could fall anywhere between the average level of student proficiency and two standard deviations above the mean. Thus, having 56% of the items on the placement examination located within the cut-point target area helps ensure a high degree of measurement accuracy for placement decisions for the different departments at the university.

Distractor analysis

A distractor analysis of the items on the placement examination found four items (i.e. 2L, 13L, 18L, and 67G) where the average proficiency level of students who correctly answered the item was lower than the average proficiency level of students who chose one of the incorrect options. For example, item 67G assesses students' knowledge of negation and its use with "would rather". The question and the options for this language use item are:

67. I would rather _________ evening classes.

a. don't take

b. not take

c. no taking

d. not taking

Table 1 shows that the students who correctly chose option b "not take" had an average proficiency level of -0.04 logits. In contrast, students who incorrectly chose option d "not taking" had a higher average proficiency level of 0.05 logits. Option d also attracted the largest percentage of students (42%). Thus, the option "not taking" is an overly attractive distractor that not only attracts a large number of students, but also attracts higher proficiency students.

Table 1. Distractor analysis of item 67G

Option	Score	Number	Percentage	Average Measure
a.	0	154	8	-0.44
b.	**1**	**592**	**31**	**-0.04**
c.	0	348	18	-0.26
d.	0	804	42	0.05
No response		263	12	-0.20

An investigation of the other three listening section items (i.e. 2L, 13L, and 18L) revealed very similar findings to that of item 67G. Items 13L and 18L had one incorrect option that attracted students with a higher average level of proficiency, while item 2L had two incorrect options that attracted students with a higher average level of proficiency (see below for details).

Examining the fit between the Rasch model and item performance

There was only one item on the placement examination that had an infit or outfit statistic that exceeded the benchmark of ±1.4. Item 2L, which is the most difficult item on the placement examination, had an outfit statistic of 1.49. This large outfit statistic indicates that there is a significant difference between the probable level of success which the Rasch model expects students to have on this item and the actually level of success students had on this item. The situation, the question, and the options for this listening item are:

Situation 2: Ken and Nancy are at a restaurant.

2. Ken _________.

a. is having steak tonight

b. stopped eating steak

c. eats steak a lot

d. prefers chicken to steak

The spoken dialogue for listening item 2L is:

NANCY: So, are you having the steak, Ken?

KEN: Actually, I'm having the chicken.

NANCY: What? I thought you really liked steak.

KEN: I do. I eat it all the time. I just don't feel like it tonight.

Table 2 shows that a group of 202 students correctly answered this item (i.e. option c). However, their average level of English proficiency is -0.18 logits lower than students who incorrectly chose either option b (0.01 logits) or option d (-0.01 logits). The large outfit statistic of 1.5 for option c suggests that there is an unexpected relationship between this very difficult item on the placement examination and students with very low levels of English proficiency.

Table 2. Distractor and fit analyses of item 2L

Option	Score	Number	Percentage	Average Measure	Outfit
a.	0	839	8	-0.20	0.9
b.	0	597	31	0.01	1.2
c.	**1**	**202**	**18**	**-0.18**	**1.5**
d.	0	507	42	-0.01	1.1
No response		16	1	-0.13	

An investigation of the 202 students' response patterns on the placement examination uncovered a group of 35 students who successfully answered item 2L despite having an average proficiency of -1.08 logits. The Rasch measurement model, which is based upon the difference between a student's level of proficiency and an item's level of difficulty, predicted that this group of low proficiency students would have a 3 percent chance of correctly answering the most difficult item on the placement examination. The discrepancy between what the Rasch model predicted and what was observed in these students' responses thus resulted in the large outfit statistic.

Discussion

Evaluating the placement examination

Overall, the placement examination performs quite well. There are very few items that are either too difficult or too easy for the students. The placement examination thus does a good job at targeting the students' level of English. Each section of the placement examination also has items located along the item difficulty continuum, which can help accurately define a wide range of student proficiency in English. There are also a good number of items on the placement examination that are located in the cut-point target area. The high degree of cut-point targeting can help ensure that the most accurate and reliable placement decisions are made.

Recommendations for improving the examination

The evaluation of the placement examination also revealed some areas that could be addressed to not only improve the performance of the examination, but also customize it to meet the specific assessment needs of the university's EFL

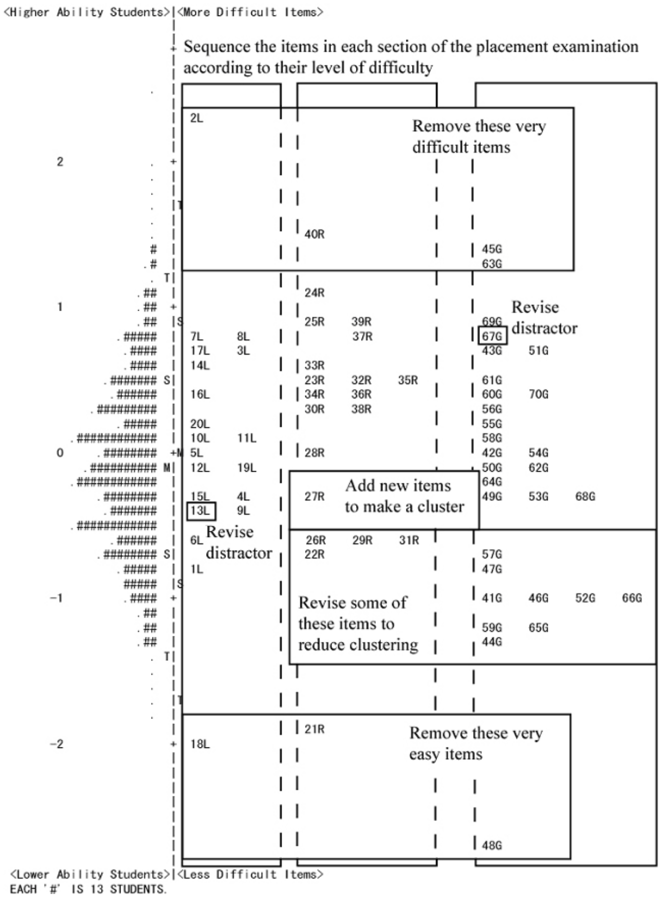


Figure 3. Suggested revisions for the placement examination

program. Figure 3 outlines some of the more immediate revisions that should be considered.

First, the items that have a level of difficulty two standard deviations either above or below the average proficiency level for the students should be removed from the placement examination. Since the purpose of the placement examination is to identify high proficiency students, very difficult or very easy items do not help achieve that goal. For example, almost every student was unsuccessful on 2L, 40R, 45G, and 63G and they were equally successful on items 21R, 18L, and 48G. As a result, these seven items do not contribute a large amount of information about students' level of English proficiency. Nevertheless, an argument could be made to retain the easiest three items on the basis that they might serve as good introductions to their respective sections on the placement examination. In other words, these items might help build students' confidence so that students do not feel overwhelmed as they take the placement examination and do not develop a negative image of the EFL program.

Second, the number of language use items clustered one standard deviation below the average proficiency level of students should either be reduced or revised in order to fill the gap that exists above item 57G. The cluster of reading items located at the bottom of the Wright map should also be revised so that they become slightly more difficult. Since the probable cut-points can be very dynamic from year to year, it is a good idea to have a clustering of items located just below the cut-point targeting area, just in case the proficiency level of new students is lower than in previous years. Adopting this strategy would thus ensure continued measurement accuracy even if the English proficiency of the student population changes from one year to the next.

Third, items that the distractor and the fit analyses have identified as being problematic should be removed or revised. Since items 2L and 18L are not contributing much information to placement decisions, these questions should be replaced with items that have a level of difficulty located around the probable cut-point. Test writers can examine other items that have this level of difficulty in order to ascertain what types of item characteristics contribute to item difficulty (see Weaver & Sato, 2008, for an example of this type of analysis). Items 67G and 13L, on the other hand, need to be slightly revised. In both cases, one of the incorrect options is attracting students who have a higher average proficiency level than the students who correctly answered the item. Test writers will thus need to determine the source of the attraction and revise the item accordingly.

Fourth, the sequence of the items on the placement examination should be examined. This commercially produced placement examination is organized in a manner in which the number of items students must answer and the amount of input they must process increase as students progress through the examination. The difficulty levels of items on the placement examination, however, do not always reflect this progression, especially in the listening section. As a result, test writers may want to consider re-sequencing items based upon their actual level of difficulty so that students first encounter easier items, followed by increasingly more difficult items.

Conclusion

The use of the three Rasch-based strategies demonstrated in this paper should be thought of as the beginning step of the evaluation and customizing process. The quantitative analyses of the items on the placement examination provide test writers with suggestions on how to improve the psychometric performance of the examination. The next step is more qualitative in nature. The focus of the investigation shifts towards determining potential reasons for unexpected item and/or distractor performance. Adopting a mixed methods approach will not only provide a deeper understanding of the placement examination, but it will also help test writers customize a commercially produced EFL placement examination so that it satisfies the specific assessment needs of their university.

During the 2007-08 academic year, **Christopher Weaver**, **Andrew Jones**, and **Juergen Bulach** conducted a series of investigations examining different aspects of placement examinations from a Rasch measurement perspective. Readers interested in how the Rasch model can be used to evaluate placement examinations and track students' level of English proficiency over time are directed to the papers listed in the reference section. Inquires can also be sent to Christopher Weaver <ctwaway@hotmail.com>.

References

Andrich, D. (1989). Distinction between assumptions and requirements in measurement in the social sciences. In J. Keats (Ed.), *Mathematical and theoretical systems* (pp. 7-16). North Holland: Elsevier Science Publishers BV.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, N. J.: L. Erlbaum.

Lesley, T., Hansen, C., & Zukowski-Faust, J. (2003). *New interchange: Passages placement and evaluation package*. Cambridge: Cambridge University Press.

Linacre, J. (2004). WINSTEPS Rasch measurement computer program (Version 3.57.1) [Computer software]. Chicago: Winsteps.com.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B. Wright ed.). Chicago: The University of Chicago Press.

Richards, J., Hull, J., Proctor, S., & Shields, C. (2006). *Full contact 1A* (3rd ed.). Cambridge: Cambridge University Press.

Weaver, C. (2008). Defining and tracking student performance on an EFL placement examination over a three-year period. *Jissen Women's University FLC Journal, 3*, 49-61.

Weaver, C., & Sato, Y. (2008). Tracking and targeting: Investigating item performance on the English section of a university entrance examination over a four year period. *JALT Journal, 30*(1), 105-128.

Wright, B., & Stone, M. (1979). *Best test design: Rasch measurement*. Chicago: Mesa Press.