Tracking and Targeting: Investigating Item Performance on the English Section of a University Entrance Examination over a 4-Year Period

Christopher Weaver
Toyo University
Yoko Sato
Tokyo University of Agriculture and Technology

This empirical study introduces population targeting and cut point targeting as a systematic approach to evaluating the performance of items in the English section of university entrance examinations. Using Rasch measurement theory, we found that the item difficulty and the types of items in a series of national university entrance examinations varied considerably over a 4-year period. However, there was progress towards improved test performance in terms of an increased number of items assessing different language skills and content areas as well as an increased number targeting test takers' knowledge of English. This study also found that productive items rather than receptive items better targeted test takers' overall knowledge of English. Moreover, productive items were more consistently located around the probable cut point for university admissions. The paper concludes with a detailed account of a number of probable factors that could influence item performance, such as the use of rating scales.

本研究は経験主義的立場に立ち、大学入試の英語の問題に用いられたテスト項目がよく機能したかどうかについて評価するための体系的アプローチとして、テスト項目が「母集団に的を絞れたか」、また「選抜ラインに的を絞れたか」いう側面に注目する方法を導入する。ある国立大学の英語の入試問題についてラッシュ分析を行った結果、テスト項目の「項目困難度」と「項目の種類」は、四年間を通してかなり多様であった。しかし、多様な言語技能と内容領域を査定する項目の増加、受験者の英語力に的を絞った項目の増加という点において、テスト項目の機能性は改善の方向に向かっていた。また受容的能力（理解力）を問う項目に比べ、産出的能力（表現力）を問う項目の方が、受験者の総合的英語力の測定により的を絞ったテスト項目になっており、さらには入学者の選抜ラインと想定されるあたりに、より集中的に配置されていた。結論では、「評価（評定）尺度」の使用等、テスト項目の機能性に影響すると考えられる要因についても詳述する。

One of the most significant challenges facing university administrators and writers of the English section of any university entrance examination is how to ensure that the difficulty of the test items does not significantly vary from year to year. Guidelines issued by the Ministry of Education, Culture, Sports, Science, and Technology along with ministry-approved textbooks inform test writers about the type of English knowledge test takers should have mastered in junior and senior high school. Yet, the actual performance of entrance examination items designed to assess test takers' level of English knowledge remains largely unexamined. Although the English section has had a long-standing role in university admission policies, there have been relatively few empirical studies (e.g., Brown & Yamashita, 1995a, 1995b; Ito, 2005; Kikuchi, 2006) investigating its performance. This paper aims not only to contribute to this important area of second language assessment, but also to introduce a systematic approach to monitoring item difficulty that takes into consideration some of the special circumstances surrounding university entrance examinations in Japan.

Strategies to Monitor Item Difficulty

There are a number of ways in which item difficulty can be monitored. Often large-scale proficiency examinations such as TOEIC® and TOEFL® use item trialing. This technique involves adding a set of items to an examination not to assess test takers' level of English knowledge, but rather to determine the level of difficulty that test takers have with these items. Unfortunately, item trialing is usually not possible with university entrance examinations because on the same day examinations are given for a number of different subjects in addition to English, which in turn limits the number of test items that can appear in any one section. Adding a set of trial items would thus seriously reduce the number of items available to determine test takers' level of ability. Moreover, many universities publicize their entrance examinations and commercial publishers sell numerous books explaining previous examinations item by item. These materials become primary study materials for many prep schools and test takers. As a result, the function of trialed items when used in a future entrance examination may be reduced to simply assessing test takers' memorization skills. The combined effect of these factors thus prevents item trialing from being a practical means of monitoring item difficulty.

Conducting a small-scale trial before the actual administration of the examination is another means of determining item difficulty. This strategem involves recruiting a group of test takers, purportedly representative of the larger test taker population, to take the examination. Their responses would then provide test writers with estimates of item difficulty so that any needed adjustments could be made before the actual administration of the examination. Test security, however, renders this scheme a virtual nonstarter for many universities.

Another technique involves using a core set of items that reoccur on two different examinations. Using Rasch measurement theory, the difficulty estimates for this common set of items would anchor the estimates of difficulty for the remaining items (Wolfe, 2000).This approach also allows test writers to examine the degree to which item difficulty varies across the different examinations. Unfortunately, many of the same challenges that prohibit the use of item trialing also prevent the reuse of a core set of items.

Targeting Item Difficulty to Test Takers' Ability Levels

Targeting is an approach that evaluates an entrance examination according to the degree to which the difficulty of the test items overlaps with the test takers' level of ability. The amount of overlap between item difficulty and test taker ability can be determined using the graphical output from a Rasch analysis, commonly referred to as a Wright map (Wright & Stone, 1979). This graphical output is valuable because test takers' level of ability and test items' level of difficulty are placed upon the same scale of reference measured in logits. In order to provide a clear explanation of targeting, a simulated data set is used to illustrate what a poorly targeted examination looks like (see Figure 1).

Considering that many readers may be unfamiliar with Wright maps, a short explanation of how to interpret this graphical output is in order. In the middle of Figure 1, there is the logit scale with its values indicated on the far left side of the figure. Once again, logits define the common scale of reference regarding the test takers and the items on the examination. For this simulated data set, the logit scale starts at -1 logits and ends at 1 logit. By itself, a logit simply indicates the relative frequency of success over the relative frequency of failure (Smith, 2000). The logit scale thus needs a point of reference to become meaningful. The meaning of Figure 1 begins with the performance of the test takers on the examination. The resulting estimates of each test taker's ability, represented with a # sign, are shown on the left side of Figure 1. Ability in the context of this investigation is the test takers' knowledge of English as defined by the items on the English section of a university entrance examination. Test takers' level of ability ranges from 0 logits to 1 logit. In other words, test takers located around 0 logits have less English knowledge than those located around 1 logit. The mean level of ability for these test takers is 0.5 logits, signified with the M marker. The S markers represent one standard distribution above and below the mean; while the T markers represent two standard distributions above and below the mean.

The right side of Figure 1 provides the second source of meaning for the logit scale. The different items on the examination, represented with a *, are placed along the scale according to their level of difficulty. For this simulated set of examination items, the range of difficulty starts at -0.08 logits and continues to 0.08 logits. Items located around -0.08 logits are less difficult, whereas items located around 0.08 logits are more difficult. The mean level of difficulty is 0 logits. Since the performance of the items is of primary interest, the standard practice is to set the starting point of the logit scale, 0 logits, at the mean for item difficulty. Once again the M, S, and T markers represent the mean for item difficulty, and the different standard distributions above and below the mean.

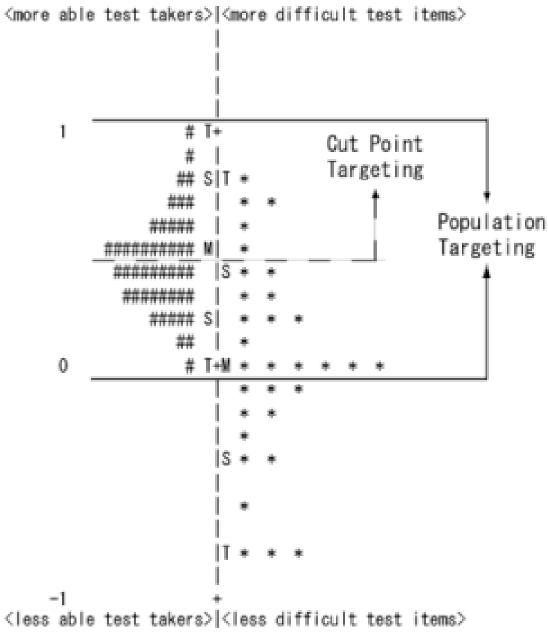

Figure 1. Wright map of a poorly targeted examination (simulated data)

Since the estimates of test takers' ability and the estimates of the item difficulty are placed upon the same scale, it is possible to compare the two directly. For example in Figure 1, a comparison of the mean for the test takers and the mean for the items produces a difference of 0.5 logits, revealing that the examination was very easy for most of the test takers. Another indicator of the ease of this examination is that the mean for item difficulty is located two standard distributions below the mean for test takers' ability. In other words, test takers with a level of English knowledge two standard distributions below the mean had a 50 percent chance of correctly answering almost half of the questions on the examination.

This imbalance between estimates of test taker ability and item difficulty does have an important implication. One's location on the logit scale is dependent upon the location of the test items. As a result, the estimate of test taker ability is more accurate when there are items in close proximity to that point on the logit scale. One of the advantages of Rasch measurement is that it provides an estimate of measurement error for every test taker and test item (Smith, 2001). Table 1 shows that for this poorly targeted examination, measurement error increases for test takers located at the higher ability levels. For example, the measurement error is four times higher for test takers located around 1 logit (0.24) than those located around 0 logits (0.06).

Table 1. Simulated test taker ability estimates and their accompanying estimates of measurement error

Test taker ability estimate	Standard error
1.00	0.24
0.86	0.20
0.78	0.19
0.68	0.17
0.58	0.15
0.40	0.12
0.34	0.11
0.22	0.09
0.13	0.08
0.00	0.06

One way to reduce the amount of measurement error is to increase the number of items that fall within the range of the test takers' ability. This concept is called targeting. By targeting the difficulty of items at the ability of the test takers, each item can provide the greatest amount of information. When test information is maximized, measurement error is minimized (Gershon, 2006). In the context of university entrance examinations, there are two types of targeting worthy of consideration. The first type involves targeting all who sit the examination. The focus here is having at least one test item located at each of the different ability levels of the test takers. This type of coverage ensures that the entire continuum of English knowledge is well defined and there is at least one item on the examination that test takers have a 50 percent chance of correctly answering. This type of coverage is called *population targeting*. Referring back to Figure 1, the population targeting is poor because too many are located at the lower levels of test takers' abilities in addition to 12 items that are below the ability level of any test taker. On the opposite end of the continuum, there are no items located around test takers who have an ability level one standard deviation above the mean. As a result, the exact location of these test takers is uncertain. Figure 2, in contrast, illustrates how items on an entrance examination can provide much better coverage of the test takers' abilities in an ideal situation.

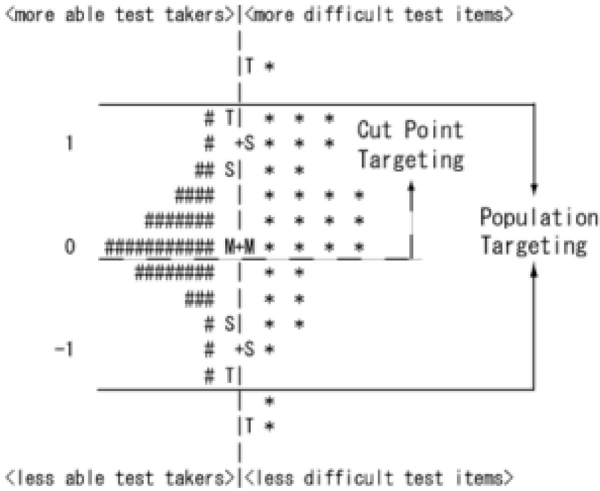

Figure 2. Wright map of a well targeted examination

The second type of targeting is of vital importance for entrance examinations. These types of tests are primarily used to make admission decisions, which in turn rely upon cut points to decide which test takers will be given an opportunity to attend the university. Such a decision becomes quite problematic when an examination has few or no test items located at the higher ability levels as shown in Figure 1. Considering the measurement error associated with poorly targeted examinations, it becomes imperative to have a group of items located around the probable cut point for admission decisions. This type of coverage is called *cut-point targeting*. The cut-point targeting in Figure 2 is a vast improvement over the population targeting, with 20 test items located around the probable cut point compared to the 5 items in Figure 1. Thus, the focus of cut-point targeting is to have a group of test items located around the probable cut point so that the items can accurately define test takers' level of English knowledge for the purpose of admission.

The ideal number or percentage of items located around the probable cut point depends upon the specific assessment needs of the university. For the purposes of the present investigation, the range of abilities where the probable cut point may fall is quite large, starting at the mean ability for test takers and extending to the most able test taker. The reason for this sizeable area is that different departments at the university have different cut points for admission decisions. As such, there is a need for a substantial number of items located around the multiple probable cut points in order to clearly define test takers' level of English knowledge.

Evaluating Targeting Over Time

Item performance across different administrations of the examination must be interpreted cautiously. Ideally, each entrance examination would feature a reoccurring set of core items that would anchor the item difficulty estimates for the other items. Since item security concerns often preclude this, population targeting and cut-point targeting are the most practical alternatives for the purposes of monitoring item difficulty. Yet, a key assumption is that the test takers' overall knowledge of English is relatively stable from year to year. This assumption, of course, is open to debate and thus should be factored in when evaluating item performance.

Research Questions

The purpose of this investigation is thus to demonstrate how population targeting and cut-point targeting can be used to monitor the difficulty of test items appearing on the English section of a university entrance examination over a 4-year period. The following research questions guided this investigation:

1. Which types of test items overlapped with test takers' knowledge of English?
2. Which types of items were located around the probable cut point for university admissions?

It is important to note that this investigation does not examine the relationship between the number of points allocated to different test items and test takers' performance. Although this is a very interesting area of research, which involves issues such as item weighting and rater effects (e.g., Myford & Wolfe, 2003), this investigation focuses upon item difficulty. As a result, it is important to clarify the relationship between item difficulty and the number of points allocated to an item. Item difficulty is defined as the proportion of incorrect responses a group of test takers have made on a particular test item. On the other hand, the number of points allocated to a particular item is a qualitative decision made (a) in advance by those who write and edit the examination; and, in certain cases, (b) afterward by those who grade the responses. In addition, there is not necessarily a direct relationship between the number of points allocated to a particular item and the level of difficulty that the item poses for test takers. While point allocations must be made before test takers sit the examination, the level of difficulty of items is not usually known until after the tests have been graded. Nevertheless, the levels of item difficulty found on previous examinations can inform decisions concerning point allocations for items to be used on future tests.

Method

A research team collected 1,996 test takers' examination responses from four consecutive administrations of the English section of an entrance examination from a national university located on the outskirts of Tokyo, Japan. This data was then submitted to a Rasch analysis implemented by WINSTEPS (Linacre, 2006) to produce the estimates of item difficulty for each examination.

Analysis

Classifying the Types of Test Items

In order to provide a more detailed account of item difficulty, the items were classified according to a) the skill or content area assessed by the test item (Brown & Yamashita, 1995a; 1995b), b) the characteristics of the test item, and c) the requirements of the test item (Gronlund, 1998). This level of detail led to initial disagreements amongst the research team concerning the exact characteristics of some test items. In these cases, the members of the research team discussed their differences until an agreed classification was reached. The following characteristics were specified:

Skills/Content Areas

The respective types of items were designed to assess the ability to:

- Translate a partial phrase or a sentence from English into Japanese.
- Describe a picture, illustration, table, or chart.
- Summarize what they have read in a reading passage.
- Comprehend or understand a reading passage or a written conversation.
- Infer meaning from a reading passage or a written conversation.
- Recognize different types of Narrative Structures such as cohesive devices.

Language of the Test Item Prompt

- Japanese or English

Language of the Test Takers' Responses (either receptively or productively)

- Japanese or English

Test Takers' Response to the Test Item

- Receptive: Test takers had to display their knowledge of English receptively (e.g., a multiple choice format).
- Productive: Test takers had to display their knowledge of

English productively (e.g., with a written response ranging from a word or a partial phrase to a paragraph).

The Source of the Test Takers' Response (Productive Items Only)

- Text-based: Test takers had to provide the correct response primarily using information appearing in an accompanying reading passage, illustration, table, and/or chart.
- Student-based: Test takers were required to provide the correct response primarily using their knowledge of English without direct assistance from an accompanying reading passage, illustration, table, or chart.

Item Format (Receptive Items Only)

- Multiple-choice: Test takers had to choose the correct response from a group of possible answers. The number of distractors accompanying the correct response is noted (i.e., CEERM2 means that this particular multiple-choice question had two distractors).
- Word bank: Test takers were required to choose a number of correct responses from a word bank of possible answers. The number of distractors accompanying the correct responses is also noted.
- True or False: Test takers were asked to indicate whether or not a statement was either true or false according to the accompanying reading passage, illustration, table, and/or chart.

This classification system produced a five-character tag for each productive test item. For example, a five-character tag of "TJJPT" refers to a test item that requires test takers to complete a Translation, the question prompt for the item was written in Japanese, test takers were required to write their response in Japanese (i.e., they translated a partial phrase or a sentence from English into Japanese), the item was Productive, and the information needed to complete the translation was found in an accompanying Text. Each receptive item has a six-character tag to account for the presence of distractors. For example, "CEERM3" refers to a test item designed to assess test takers' Comprehension skills, the question prompt for the item was written in English, the possible answers were written in English, the item was Receptive, the item format was Multiple-choice, and there were 3 distractors along with the correct response.

Results

The 4-Year Overall Performance

Table 2 shows the overall performance over a 4-year period, defined in terms of a) the number of items used in each examination, b) the number and the percentage of items that overlapped with test takers' level of ability (i.e., population targeting), and c) the number and the percentage of items that had a level of difficulty located around the probable cut point for university admissions (i.e., cut-point targeting). This information is also provided for items designed to assess test takers' productive and receptive knowledge of English.

Table 2. The overall performance of the items on the English section of the university entrance examination over a 4-year period

Items	Year 1		Year 2		Year 3		Year 4	
Total number of items	23		39		40		36	
Population targeted	16	70%	17	44%	30	75%	34	94%
Cut-point targeted	10	43%	9	23%	19	48%	12	33%
Productive items	12	52%	26	67%	24	60%	23	64%
Population targeted	11	92%	15	58%	15	63%	23	100%
Cut-point targeted	7	58%	8	31%	11	46%	9	39%
Receptive items	11	48%	13	33%	16	40%	13	36%
Population targeted	5	45%	2	15%	15	94%	11	85%
Cut-point targeted	3	27%	1	8%	8	50%	3	23%

The number of items on the English section ranged from the mid-30s to 40. The exception was Year 1 with 23 items. The percentage of items that targeted the ability level of the test taker population varied considerably: the highest was 94% in Year 4, the lowest was 44% in Year 2. In terms of items located around the probable cut point, the percentage varied from 48% in Year 3 to 23% in Year 2.

Over the period of four years, productive items composed over 50% of the items on the examination. Year 2 had the highest percentage of productive items with 67%. The percentage of productive items targeting the test taker population varied considerably, from 100% in Year 4 to 50%

in Year 3. Finally, the percentage of productive items located around the probable cut point was highest in Year 1 (58%).

Similarly, the receptive-knowledge items varied in terms of population targeting. The highest percentage occurred in Year 3 (94%). Year 3 also had the highest percentage of receptive items located around the probable cut point with 50%. The percentage of receptive items targeting the probable cut point in the other years was, however, significantly lower.

The Performance of the Productive Measures of Test Takers' English Knowledge

Table 3 shows the frequency of the different types of productive items used over a 4-year period. The frequency of these items' level of difficulty overlapping with test takers' knowledge of English (i.e., population targeting) and the frequency of these items' level of difficulty being located around the probable cut point (i.e., cut-point targeting) are also shown.

Over the 4-year period, there were a number of different types of productive items utilized. The most commonly occurring item types were TJJPT, IJJPT, SJJPS, and CEEPT. Generally, each examination featured a group of item types that composed the majority of productive items. In Year 1 the combination of SEEPT, TJJPT, SJJPS, and CJJPT items composed 76% of the productive items; in Year 2 CEEPT and CEEPS items combined for 69%; in Year 3 NEEPS and DEEPS items reached 71%, and in Year 4 CEEPT, CEEPS, and CJEPT items combined for 72% of the productive items.

In terms of targeting, the level of difficulty for the different types of productive items largely overlapped with the test takers' knowledge of English. The only exceptions were the SEEPT items in Year 2 and the SEEPS item in Year 3. The percentage of productive items located around the probable cut point was generally lower. Each examination had at least one type of productive item with a level of difficulty not located around the probable cut point: TJJPT, IJJPS, and IEEPT in Year 1; TJJPT, IJEPT, and SEEPT in Year 2; SEEPS and CEEPT in Year 3, and CJEPT in Year 4.

Table 4 shows the collective performance of the different types of productive items as well as their level of difficulty compared to the location of the probable cut point. The majority of item types, which were not located around the probable cut point, had a level of difficulty that was lower than the average ability of the test takers. In other words, these items did not pose a significant challenge for the test takers. The only

Table 3. The performance of productive items over 4 years

Items	Year 1						Year 2					
	Occurred		Pop. Targeted		Cut Point Targeted		Occurred		Pop. Targeted		Cut Point Targeted	
TJJPT	2	17%	1	50%	0	0%	3	12%	3	100%	0	0%
IJJPS	1	8%	1	100%	0	0%						
IJJPT	1	8%	1	100%	1	100%	1	4%	1	100%	1	100%
SJJPS	2	17%	2	100%	2	100%						
CJJPT	2	17%	2	100%	2	100%						
IJEPT							1	4%	1	100%	0	0%
CJEPT												
IEEPT	1	8%	1	100%	0	0%						
SEEPT	3	25%	3	100%	2	67%	3	12%	0	0%	0	0%
SEEPS												
CEEPT							8	31%	3	38%	3	38%
CEEPS							10	38%	7	70%	4	40%
NEEPS												
DEEPS												
Total	12	100%	11	92%	7	58%	26	100%	15	58%	8	31%

Items	Year 3						Year 4					
	Occurred		Pop. Targeted		Cut Point Targeted		Occurred		Pop. Targeted		Cut Point Targeted	
TJJPT	2	8%	2	100%	2	100%	1	4%	1	100%	1	100%
IJJPS												
IJJPT							1	4%	1	100%	1	100%
SJJPS	1	4%	1	100%	1	100%	2	9%	2	100%	1	50%
CJJPT												
IJEPT							2	9%	2	100%	2	100%
CJEPT							4	17%	4	100%	0	0%
IEEPT												
SEEPT												
SEEPS	1	4%	0	0%	0	0%						
CEEPT	3	13%	2	67%	0	0%	9	39%	9	100%	3	33%
CEEPS							4	17%	4	100%	1	25%
NEEPS	9	38%	3	33%	3	33%						
DEEPS	8	33%	7	88%	5	63%						
Total	24	100%	15	63%	11	46%	23	100%	23	100%	9	39%

two exceptions were four NEEPS items and one DEEPS item which had a level of difficulty that surpassed the ability of test takers who were located two standard deviations above the average test taker. In short, these items were quite difficult.

Table 4. The collective performance of the productive items

Items	Occurred	Cut-point targeted		Below cut point		Above cut point	
TJJPT	8	3	38%	5	63%	0	0%
IJJPS	1	0	0%	1	100%	0	0%
IJJPT	3	3	100%	0	0%	0	0%
SJJPS	5	4	80%	1	20%	0	0%
CJJPT	2	2	100%	0	0%	0	0%
IJEPT	3	2	67%	1	33%	0	0%
CJEPT	4	0	0%	4	100%	0	0%
IEEPT	1	0	0%	1	100%	0	0%
SEEPT	6	2	33%	4	67%	0	0%
SEEPS	1	0	0%	1	100%	0	0%
CEEPT	20	6	30%	14	70%	0	0%
CEEPS	14	5	36%	9	64%	0	0%
NEEPS	9	3	33%	2	22%	4	44%
DEEPS	8	5	63%	2	25%	1	13%
Total	85	35		45		5	

The Performance of the Receptive Measures of Test Takers' English Knowledge

Table 5 shows the frequency of the different types of receptive items used in the English section of the university entrance examination over a 4-year period. The frequency of these items' level of difficulty overlapping with test takers' knowledge of English and the frequency of these items' level of difficulty being located around the probable cut point are also shown.

Table 5. The performance of the receptive measures of test takers' English knowledge

	Year 1						Year 2					
Items	Occurred		Pop. Targeted		Cut Point Targeted		Occurred		Pop. Targeted		Cut Point Targeted	
SJERW0												
CEERW0							4	31%	0	0%	0	0%
CEERW1							6	46%	0	0%	0	0%
NJERW2												
IJERW3												
CJERW4												
CEERM2	6	55%	3	50%	2	33%						
CEERM3	5	45%	2	40%	1	20%	3	23%	2	67%	1	33%
CJERM3												
CJERTF												
Total	11	100%	5	45%	3	27%	13	100%	2	15%	1	8%

	Year 3						Year 4					
Items	Occurred		Pop. Targeted		Cut Point Targeted		Occurred		Pop. Targeted		Cut Point Targeted	
SJERW0							5	38%	5	100%	0	0%
CEERW0												
CEERW1												
NJERW2	5	31%	5	100%	1	20%						
IJERW3	2	13%	2	100%	1	50%						
CJERW4	2	13%	1	50%	1	50%	2	15%	2	100%	2	100%
CEERM2												
CEERM3							6	46%	4	67%	1	17%
CJERM3	2	13%	2	100%	2	100%						
CJERTF	5	31%	5	100%	3	60%						
Total	16	100%	15	94%	8	50%	13	100%	11	85%	3	23%

The 4-year period had two distinct patterns. During the first 2 years, the examinations items exclusively assessed test takers' comprehension skills. The next 2 years, however, featured a greater variety of receptive items that assessed other skills and content areas such as summarize, inference, and narrative structures in addition to test takers' level of reading comprehension. The number of receptive items was generally stable over the 4-year period with the exception of Year 3 with 16 receptive items, which coincides with a greater range of skills being assessed.

In terms of population targeting, the receptive items varied considerably over the 4-year period. Year 2 had the poorest coverage with only 2 out of 13 items located within the test takers' overall level of English knowledge, which was a significant drop from 5 out of 11 items in Year 1. Years 3 and 4 performed much better with only one receptive item in Year 3 and two items in Year 4 not targeting the test takers' overall English knowledge.

The percentage of receptive items located around the probable cut point also varied considerably over the 4-year period. Year 3 had the highest percentage with 53% followed by Year 1 (27%) and Year 4 (17%). Not surprisingly, Year 2 had the lowest percentage with only 8% of the receptive items located around the probable cut point. Table 6 shows that all of the receptive items not located around the probable cut point had a level of difficulty lower than the average ability level of the test takers. The only exception was two difficult CJERTF items in Year 3.

Table 6. The collective performance of the receptive items

Items	Occurred	Cut-point targeted		Below cut point	
SJERW0	5	0	0%	5	100%
CEERW0	4	0	0%	4	100%
CEERW1	6	0	0%	6	100%
NJERW2	5	1	20%	4	80%
IJERW3	2	1	50%	1	50%
CJERW4	4	3	75%	1	25%
CEERM2	6	2	33%	4	67%
CEERM3	14	3	21%	11	79%
CJERM3	2	2	100%		
CJERTF	5	3	60%	2	40%
Total	53	15		38	

Table 7 shows the performance of the different item formats (i.e., Multiple-choice, Word bank, and True or False items) as well as their performance according to the number of distractors. During the 4-year period, the receptive items were predominantly multiple choice items (22) or word bank items (26). These two item formats performed similarly in terms of population targeting with 59% of multiple choice items and 58% of word bank items targeting the test takers' overall knowledge of English. These two item formats, however, differed in terms of the percentage of items located around the probable cut point. Multiple choice items had 32% cut-point targeting compared to 19% for word bank items. Although True or False items were used only in Year 3, they performed quite well with 100% population targeting and 60% cut-point targeting.

Table 7. The performance of different item formats and their performance according to the number of distractors

Item Formats	Occurred	Population targeted		Cut-point targeted	
M	22	13	59%	7	32%
W	26	15	58%	5	19%
TF	5	5	100%	3	60%
W0	9	5	56%	0	0%
W1	6	0	0%	0	0%
W2	5	5	100%	1	20%
W3	2	2	100%	1	50%
W4	4	3	75%	3	75%
M2	6	3	50%	2	33%
M3	16	10	63%	5	31%

The use and performance of distractors in multiple choice and word bank items varied considerably during the 4-year period. Whereas the multiple choice questions had either two or three distractors the word bank items ranged from no distractors to four. In terms of population targeting, three distractors performed better than two for multiple choice items. For the word bank items, having no distractors or only one distractor resulted in poorer performances. In terms of the percentage of multiple choice format items targeting the probable cut point, two distractors performed better than having three distractors. Word bank items, on the other hand, had better cut-point targeting with an increased numbers of distractors.

Discussion

Once again, the implications arising from the results must be considered carefully since the item difficulties over the four years are not anchored to a common set of items. The discussion then examines a number of factors that might underlie the performance of the productive and receptive items on the different examinations. This study focuses upon the characteristics of the different item types and does not take into consideration linguistic factors, such as vocabulary level or the level of readability which may also mediate the interaction between the test takers and the examination (see Weaver & Sato, 2008, for an example of this type of analysis).

Overall Performance of the English Section of the University Entrance Examination Over a 4-Year Period

This investigation reveals a considerable amount of variation from year to year. For example, Table 2 shows that the number of items is almost twofold between Year 1 and Years 2, 3, and 4. The initial increase of test items, however, did not necessarily improve performance. This finding is counter to conventional thinking that an increased number of items leads to improved test performance in terms of reliability (Traub & Rowley, 1991). Although the correlation-based reliability coefficient of the entrance examination increased by 0.04 from Year 1 to Year 2, the percent of population-targeted items fell from 70% in Year 1 to 44% in Year 2 despite an increase of 16 items. The additional items in Year 2 also did not help increase the number of items located around the probable cut point. In Years 3 and 4, the correlation-based reliability coefficient continued to increase by 0.05 each year. In addition, the percentage of items targeting the test taker population continued to increase to 75% in Year 3 and 94% in Year 4.

In terms of the percentage of items located around the probable cut point, Year 3 (48%) exceeded the level reached in Year 1 (43%), doing so in two distinctive ways. In Year 1, the productive items performed better than the receptive items. In Year 3, the performances of the productive and receptive items were more balanced.

Potential Factors Underlying the Performance of the Productive Items

Over the 4-year period, reading comprehension was the most commonly tested skill with 31 out of 40 items targeting the test takers' overall knowledge of English and 13 items located around the probable cut point (see Table 3). A factor that had a consistent influence on this type of item was whether or not test takers were required to respond in Japanese or English. The Japanese-response items (i.e., the two CJJPT items in Year 1) were more difficult than the items requiring responses in English (i.e., the four CJEPT items in Year 4). Table 3 shows that the CJJPT items were located around the probable cut point; in contrast, the CJEPT items were located below the mean ability level of the test takers, but still within the population target. One possible explanation for this difference is that requiring test takers to demonstrate their level of reading comprehension productively in English may be a relatively easy task since it requires test takers to identify what needs to be comprehended in reading text and transfer this information to their answer sheet. CJJPT items, on the other hand, require the additional steps of translating the information from the reading passage into Japanese as well as summarizing and synthesizing information from the reading passage. Another source of support for this explanation is a study that found that higher levels of cognitive load generally led to increased levels of item difficulty for reading comprehension questions used on a university entrance examination (Weaver & Romanko, 2005).

An interesting extension to this finding is the comparison between productive items with question prompts written in Japanese that required Japanese responses from test takers versus items with English question prompts requiring English responses. Table 3 shows that although English prompt/response items (59) occurred almost three times as often as Japanese prompt/response items (19) during the 4-year period, Japanese prompt/response items performed at a higher level. In terms of population targeting, 95% of the Japanese prompt/response items targeted test takers' overall level of English knowledge compared to 66% of the English prompt/response items. The difference between these two types of items also was apparent with cut-point targeting: 63% of the Japanese prompt/response items compared to 36% of the English prompt/response items. However, this finding should not be used as a justification for the use of Japanese prompt/response items. Rather this finding highlights a unique challenge that faces foreign-language-test writers. Table 4 shows that the majority of English prompt/response items were located below the mean ability level for the test takers and thus were within the realm of their English knowledge. As a result, test writers need to design items that require more than identification and copying skills from test takers. However, Table 4 also shows that the difficulty level for one DEEPS item and four NEEPS items was beyond the ability level of the most able test taker in Year 3. In other words, these items designed to assess test takers' descriptive skills and knowledge of narrative structures were far too difficult and thus reveal the challenge of writing English prompt/response items located around the probable cut point.

Another interesting finding is that the productive items located around the cut point assessed a number of different skills and content areas over the 4-year period. Such variety not only helps to create a more comprehensive account of English knowledge, but also lends support to the argument that the examinations evaluate more than test takers' grammatical competence (e.g., Guest, 2000). It is hoped that this finding will have a positive washback effect on future test takers and their teachers: that a well rounded knowledge of English is important.

Potential Factors Underlying the Performance of the Receptive Items

The performance of the receptive items reveals an important rationale for tracking and targeting items. Table 5 shows that the receptive items in Years 1 and 2 focused exclusively on reading comprehension skills using English question prompts and English response choices. These items unfortunately did not provide significant amounts of information about the test takers' overall level of English knowledge, especially in Year 2 with only 15% of the items falling within the population target. Years 3 and 4, however, featured receptive items that assessed a larger range of skills and content areas and utilized a variety of question prompt/response choice formats. These changes resulted in an increased number of receptive items targeting the test takers' overall level of English knowledge. Year 3 also had the highest percentage of receptive items (50%) located around the probable cut point. The introduction of new types of receptive items, however, must be considered as a work in progress. For example, the five NJERW2 items in Year 3 successfully targeted the test-taking population, but had only one item located around the probable cut point. The five SJERW0 items in Year 4 also had a similar performance with good population targeting, but poor cut-point targeting. A systematic approach of tracking and targeting can provide test writers with vital information about how new types of receptive items performed in order to maintain or improve their performance in future entrance examinations.

Similar to the productive items, there were a number of factors that influenced the performance of receptive items. Although multiple choice and word bank items had similar amounts of success targeting the population of test takers, a greater percentage of multiple choice items were located around the probable cut point. This finding highlights an important design feature that differentiates these two types of item formats: whereas the possible answers for a multiple choice item are exclusive to one item, a number of different receptive items can share a common word bank. One implication of a shared word bank is that the number of possible answers decreases as test takers complete the different items. As a result, items that initially have a level of difficulty located around the probable cut point may become easier through a process of elimination. A means of circumventing this shortcoming is to design items so that possible answers can be used more than once. However, designing items so that alternative answers are a credible choice for multiple items can be a formidable challenge.

Another factor that influenced the performance of receptive items was the number of distractors accompanying the correct response. The influence of this factor, however, varied according to the item format. The number of distractors had a relatively consistent effect upon the performance of word bank items. Generally, an increased number of distractors led to higher percentages of population and cut-point targeting. An increased number of distractors in multiple choice items, on the other hand, resulted in better population targeting but poorer cut-point targeting. This finding provides partial support for the Shizuka, Takeuchi, Yashima, & Yoshizawa (2006) suggestion that traditional four-option multiple choice items can be reduced to three alternatives without sacrificing test performance.

Overall we found that relatively few receptive items were located around the probable cut point with the exception of Year 3. During the 4-year period, only 15 out of 53 receptive items reached this level of difficulty. This stands in contrast to the productive items. Productive items such as translations or written compositions usually utilize polychotomous rating scales. As a result, the productive items employing a multiple-point rating scale can provide partial credit for test takers' responses and thus define a larger range of English knowledge. For example, Weaver and Sato (2007) found that a 3-point rating scale was optimal for assessing test takers' grammatical competence set within a communicative situation. Receptive items, in contrast, score test takers' responses as either right or wrong. Dichotomous rating scales thus define a very specific level of English knowledge. It is possible to have receptive questions that utilize a polychotomous rating scale where test takers receive credit for a response choice that is not entirely correct, but reveals attainment of some developmental stage on the way towards target-like use. Designing this type of receptive item, however, requires a great deal of planning and care. Moreover, test writers will need to be versed in processability theory (Pienemann, 1998) and the research concerning the developmental steps of different grammatical features such as negation (e.g., Batet & Grau, 1995), wh-question formation (e.g., Mackey, 1999), and relative clauses (e.g., Diessel, 2004).

Conclusion

The English section of a university entrance examination provides test writers with a multitude of challenges. In many cases, a strictly defined time limit combined with historical influences of how things are done govern the number and the types of items that appear. As such, a systematic approach focusing upon the performance of previous test items can provide test writers with an essential source of information. Tracking and targeting items allows test writers to gain a deeper understanding of how different factors potentially mediate item performance. Since most entrance examinations do not share a common set of items, test writers should be cautious when comparing different test performances over time. In other words, they should be continually on the outlook for consistent trends that appear over multiple administrations of the examination. Another focal point for test writers should be the importance of cut-point targeting in order to ensure the highest degree of measurement accuracy. In essence, the whole idea is to transform hindsight gained from previous item performances into foresight which can help improve future performance.

Acknowledgments

*The authors would like to acknowledge the past and the present governing body of Tokyo University of Agriculture and Technology for their support of this research project. In addition, gratitude is extended to the many people who have contributed to this research in one way or another. This investigation was funded by a JSPS grant-in-aid of research ([*平成17年度〜平成18年度科学研究費補助金基盤研究(C) 課題番号17520371*]).*

Christopher Weaver is a lecturer at Toyo University. His area of research includes task-based instruction, individual differences, and psychometrics with a special focus on practical applications of Rasch measurement theory.

Yoko Sato is a Professor of English at Tokyo University of Agriculture and Technology. Her main research interest focuses upon a close textual analysis of poetry and drama, language testing, and vocabulary learning strategies. She is a cotranslator of Morton N. Cohen's *Lewis Carroll: A Biography* (1999).

References

Batet, M., & Grau, M. (1995). The acquisition of negation in English. *Atlantis, 17* (1), 27-44.

Brown, J. D., & Yamashita, S. (1995a). English language entrance examinations at Japanese universities: 1993 and 1994. In J. D. Brown & S. Yamashita (Eds.), *Language teaching in Japan* (pp. 86-100). Tokyo: JALT.

Brown, J. D., & Yamashita, S. (1995b). English language tests at Japanese universities: What do we know about them? *JALT Journal, 17* (1), 7-30.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge: CUP.

Gershon, R. (2006). Understanding Rasch measurement: Computer adaptive testing. *Journal of Applied Measurement, 6* (1), 109-127.

Gronlund, N. (1998). *Assessment of student achievement* (6th ed.). Boston: Allyn and Bacon.

Guest, M. (2000). "But I have to teach grammar!": An analysis of the role "grammar" plays in Japanese university English entrance examinations. *The Language Teacher, 20* (11), 23-29.

Ito, A. (2005). A validation study on the English language test in a Japanese national wide university entrance examination. *Asian EFL Journal, 7* (2), 91-117.

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities a decade later. *JALT Journal, 28* (1), 77-96.

Linacre, J. (2006). *WINSTEPS Rasch measurement computer program* (Version *3.60*). Chicago: Winsteps.com.

Mackey, A. (1999). Input, interaction and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition, 21* (4), 557-587.

Myford, C., & Wolfe, E. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4* (4), 386-422.

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam: Benjamins.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three- and four-point English tests for university entrance selection purposes in Japan. *Language Testing, 23* (1), 35-57.

Smith, E.(2000). Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement, 1* (3), 303-326.

Smith, E., (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2* (3), 281-311.

Traub, R., & Rowley, G. (1991). Understanding reliability. *Educational Measurement: Issues and practice, National Council on Measurement in Education, 10* (1), 37-45.

Weaver, C., & Romanko, R. (2005). Assessing oral communicative competence in a university entrance examination. *The Language Teacher, 29* (1), 3-9.

Weaver, C., & Sato, Y. (2007). Jukensha no eigo-communication-noryoku-shikibetsu no tameno hyouka-shakudo ni tsuite–bunseki-rei–Rasch model ga kitaisuru hyouka-shakudo no performance [The use of rating scales to differentiate test takers' English communicative competence: the Rasch model expectations of a well performing rating scale]. *Daigaku Nyushi Kenkyu Journal*, No.17, 135-142.

Weaver, C., & Sato, Y. (2008). Estimating and determining the difficulty of English reading passages used in a university entrance examination. Manuscript submitted for publication.

Wolfe, E. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement, 1* (4), 409-434.

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