The Development of an Item Bank for English Testing and a Computer Program for Estimating Ability

Chisato SAIDA
Namiki Senior High School, Ibaraki Prefecture
Tamaki HATTORI
University of Tsukuba, Institute of Psychology

Abstract

An item bank is a collection of test items that have been calibrated onto a common scale. The item response theory makes it easy to develop item banks. An item bank can be used in creating not only classroom English tests but also large-scale English tests. The development of tests with the assistance of an item bank enables teachers and test developers to save the time and energy they would have to spend on conventional test development. The biggest advantage of using an item bank is that it becomes possible to compare the results of different tests. This will improve assessment based on data analysis in education. Recently, in Japan, private testing agencies have developed English proficiency tests using their own item banks. However, very few item banks have been made available for schoolteachers to use on a regular basis. This article explains the development of an item bank for practical use in the classroom and in large-scale program assessment for high school English education. In addition, a user-friendly computer program was developed to estimate students' scores from item bank testing.

1. Introduction

Computer technology has had an effect not only on teaching methods, such as the computer assisted language learning (CALL) system, but also on assessment methods as shown in the development of computerized adaptive testing (CAT). Item response theory models and item banking facilitate the introduction of computer technology in educational assessment (Embretson & Reise, 2000). An item bank is a collection of test items in which statistical characteristics (e.g., item difficulty parameters and item discrimination parameters) are calibrated on a common scale. Item response theory (IRT) is very useful in developing
item banks, because IRT makes it easier to link item parameters to a common scale (Hambleton & Swaminathan, 1985). The use of an item bank for creating tests makes it possible for teachers and test developers to improve the quality of testing and assessment as well as to reduce the effort and expense of preparing new items every time the test is administered.

The development of tests using item banks has been in practice in the USA since 1990 (Thissen, 2001). TOEFL, a test of English as a foreign language, is a well-known English test using a huge item bank produced with the application of item response models (http://www.toefl.org/research/reports.html). A computer-based TOEFL (CBT) was implemented in 1998. Item banks, however, have been used for both large-scale proficiency tests and small-scale achievement tests. Any test administrator can benefit from using item banks to create tests (Ward & Murray-Ward, 1994). In the USA, item banks are purchased or leased along with the computer programs to operate them.

In Japan, on the other hand, the advantages of using item banks in the development of tests have primarily been recognized by private testing agencies. These testing agencies have developed English proficiency tests for Japanese students using item response models and item banks. Examples include GTEC by Benesse Corporation (1998), CASEC by the Society of Testing English Proficiency (2001), TOEIC Bridge by the International Business Communication Association (2001), and ACT by Kirihara Shoten (2002). However, there are very few item banks available for schoolteachers to create classroom English tests or large-scale tests. The development of an item bank, in general, costs a great deal of money and takes a large amount of time, because a large number of people (item writers, item analysts or psychometric specialists, test administrators) are needed in the process. Thus, item banks developed by private testing agencies are not available to the public. Nakamura (1998) tried to produce an item bank for English listening tests for high school students. Saida (2003a) tried to produce an item bank for classroom English tests. Both were trying to make item banks that could be used by high school English teachers. However, the number of items was too small for practical use. In addition, the application of the Rasch model limited their item banks to only estimating item difficulty parameters. To enhance the function of item banks, the application of a more appropriate item response model must be considered.

This research was undertaken (1) to develop an item bank for schoolteachers to create regular classroom tests or large-scale high school English tests, and (2) to develop a computer program to estimate ability parameters, which are necessary to identify after the administration of item bank testing.
2. Method

2.1 Item Bank Development

In general, item banks are developed in four stages (Hambleton, 1989).

First Stage: Item Writing

Items are created according to the specifications of the test. Sufficient numbers of content-valid items should be created to enable the combining of items in various ways. Some items will be discarded due to statistical characteristics when developing the item banks.

Second Stage: Administration of Pre-tests

The items are divided into several test forms to administer to a number of test takers. Each form generally contains anchor items in order to make it possible to compare item parameters in each form through equating. This technique is called the anchor test design, and it is often employed in the development of item banks.

Third Stage: Item Parameter Estimation

Item parameters are estimated using an appropriate item response model. Items with extremely high or low difficulty levels and/or low discrimination levels are discarded. When the scores are dichotomous (the answer is either right or wrong), a one-, two-, or three-parameter logistic item response model will be chosen, based on the number of test takers taking each variation of the test and other factors. The one-parameter logistic model known as the Rasch model is appropriate when the number of test takers is more than one hundred. The Rasch model only gives estimates for item difficulty parameters. In the language testing field, this model is used frequently because of the number of test takers. The two-parameter logistic model is appropriate for approximately five hundred test takers. This model estimates item discrimination parameters as well as item difficulty parameters. The three-parameter logistic model is appropriate when each form has around one thousand test takers. This model estimates guessing parameters in addition to item difficulty and discrimination parameters. The higher the number of item parameters to estimate, the more complicated the handling becomes (Hambleton & Swaminathan, 1985).

Fourth Stage: Equating Item Parameters

Finally, the item parameters which were estimated on the different forms are linked together to a common scale. They are then compared using the statistical information of the anchor items in each form. Items are stored in the item bank with the information of contents and item parameters.

In this study, the basic procedure outlined above was followed, with two notable changes. The first unique point of this study is in the first stage. The first stage is the most time-consuming. To avoid this problem, this study used items which had been made by high school teachers and used in the past. High school English teachers in Ibaraki prefecture
developed the Ibaraki Prefecture-Wide English Test for high school students (IPET). This study recycled items of the IPET administered from 1995 to 2002. Items consisted of four sections: listening comprehension, vocabulary knowledge, grammatical knowledge, and reading comprehension covering all English learning domains. They were divided into eight parts (see details in Saida, 2002). These items had already been administered to more than 10,000 test takers each year, so there could be no doubt that the most accurate estimates of item parameters were obtained. The use of recycled items allowed this study to dispense with the second stage, pre-testing, because the item parameters have already been estimated. Considering the number of test takers and the desired user-friendliness of the item bank, the two-parameter logistic model was chosen for the item parameter estimation. The item bank was thus provided with the information of item difficulty and item discrimination. Item parameters were estimated using BILOG (Mislevy & Bock, 1990) after removing items which did not meet the unidimensionality assumption.

The second point of diversion is in the fourth stage. The tests used for this study did not contain anchor items. Therefore, to equate item parameters, an additional seven tests were created with a combination of items drawn from the 1999 test and one of the other years. These tests were administered to 3,155 first year high school students to calculate the equating coefficients, which transformed item parameters to a common scale. The procedure of equating was detailed in Saida (2003b) and Saida and Hattori (2003; 2004).

2.2 Computer Program Development

It was necessary to develop a computer program to estimate ability scores resulting from the administration of tests constructed from item banks. Teachers and test developers can construct tests with any combination of items drawn from an item bank according to the purpose of the test. Student ability can be estimated using item parameter estimates on a common scale, unlike in classical testing theory. BILOG, the most common computer program used in item response theory, can be used for the estimation of ability parameters. However, BILOG requires a more complicated process to estimate ability parameters given item parameter estimates in this situation. In addition, as BILOG is not in general use amongst teachers, most teachers do not have access to it. Therefore, a more user-friendly computer program was developed. This program is designed to estimate ability parameters when item parameter estimates are given. The computer program, which is called EST 3, was developed using Fujitsu FORTRAN V3.0. It can be used on Microsoft Windows 98 or Windows XP.
3. Results

3.1 Item Bank

In total, 316 items were stored in the item bank. Item parameters were estimated on a common scale. The items were divided into four sections. The descriptive statistics are shown in Table 1. The variable $a$ indicates discrimination parameters and $b$ is the item difficulty parameters. Mean and standard deviation of ability parameters estimated from the results of about 140,000 test takers were set at 0.0 and 1.0 respectively. The mean and standard deviation of the item discrimination parameter estimate is 0.91 and 0.41 respectively, which indicates this item bank contains a lot of high quality items. The mean difficulty parameter estimate of the 316 items is 0.14, and the standard deviation is 1.28. The difficulty parameter estimates range from -3.12 at the minimum to 4.31 at the maximum. Therefore, this item bank can be used for a wide range of student ability. Examples of items and item parameters are shown in Appendix 1.

Table 1. Descriptive Statistics of the Item Bank

<table>
<thead>
<tr>
<th>Section Part</th>
<th>Listening</th>
<th>Phonics</th>
<th>Vocabulary</th>
<th>Idiom</th>
<th>Voc.</th>
<th>Grammar</th>
<th>Dialogue</th>
<th>Grammar</th>
<th>Sentence</th>
<th>Reading</th>
<th>Reading</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.of items</td>
<td>33</td>
<td>34</td>
<td>36</td>
<td>28</td>
<td></td>
<td></td>
<td>35</td>
<td>57</td>
<td>34</td>
<td>59</td>
<td>1</td>
<td>316</td>
</tr>
<tr>
<td>$a$ mean</td>
<td>0.75</td>
<td>0.84</td>
<td>1.02</td>
<td>0.91</td>
<td>0.99</td>
<td>0.98</td>
<td>0.94</td>
<td>0.94</td>
<td>0.81</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.D.</td>
<td>0.20</td>
<td>0.36</td>
<td>0.41</td>
<td>0.46</td>
<td>0.52</td>
<td>0.44</td>
<td>0.46</td>
<td>0.46</td>
<td>0.34</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min.</td>
<td>0.39</td>
<td>0.39</td>
<td>0.49</td>
<td>0.37</td>
<td>0.33</td>
<td>0.31</td>
<td>0.36</td>
<td>0.36</td>
<td>0.29</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td>1.22</td>
<td>1.72</td>
<td>1.95</td>
<td>2.36</td>
<td>2.36</td>
<td>1.77</td>
<td>2.21</td>
<td>2.21</td>
<td>1.50</td>
<td>2.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b$ mean</td>
<td>-0.25</td>
<td>0.02</td>
<td>-0.50</td>
<td>0.45</td>
<td>0.11</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
<td>0.58</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.D.</td>
<td>1.62</td>
<td>1.39</td>
<td>0.91</td>
<td>1.23</td>
<td>1.44</td>
<td>1.37</td>
<td>0.84</td>
<td>1.05</td>
<td>1.69</td>
<td>-3.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min.</td>
<td>-3.12</td>
<td>-1.77</td>
<td>-1.83</td>
<td>-1.52</td>
<td>-1.74</td>
<td>-1.91</td>
<td>-1.37</td>
<td>-1.69</td>
<td>-3.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td>3.02</td>
<td>3.63</td>
<td>1.50</td>
<td>2.73</td>
<td>4.28</td>
<td>4.31</td>
<td>2.52</td>
<td>2.90</td>
<td>4.31</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2 Standard Error of Ability Parameter

The maximum likelihood (ML) estimator of the ability parameter has a normal asymptotic distribution with mean $\theta$ and variance $1/I(\theta)$, where $\theta$ is true ability level and $I(\theta)$ is the amount of test information. The square root of $1/I(\theta)$ is the standard error of the ML estimator. The greater the amount of test information at a given ability level, the more closely the ML estimator of ability cluster around the true ability level, hence, the more precise the estimate (Baker, 1992). The amount of test information of the item bank produced in this study is graphed in Figure 1 and the standard error of the ML estimate is in Figure 2.
Figure 1. The amount of information of the item bank

Figure 2. The standard error of the ML estimate

When constructing a high school test using items from the item bank, forty to fifty items are generally used on one form. If fifty items are selected from the item bank at random, the amount of information and the standard error of the ML estimate can be calculated based on those fifty items. These are also graphed in Figures 1 and 2. For example, at the three points $\theta = -1.0, 0.0, 1.0$, the standard errors of the ML estimate for a fifty-item test are equal to about 0.31, 0.33, 0.42.

The confidence interval of the ML estimate is calculated with the following formulas.

Lower limit = ML estimate of ability parameter - $z \times$ standard error \hspace{5pt} (1)
Upper limit = ML estimate of ability parameter + $z \times$ standard error \hspace{5pt} (2)

In these formulas, $z$ is a normal deviate of standard normal distribution for a proper confidence interval. The $z$ is 1.64 for a ninety percent confidence interval, and 1.96 for ninety-five. Then, the ninety percent confidence interval of the ML estimate of -1.0 is between -1.51 and -0.49, 0.0 is between -0.54 and 0.54, and 1.0 is between 0.31 and 1.67. The mean ability parameter of test takers in 1999 is set at 0.0 and the standard deviation is set at 1.0. Therefore the Z-score of -1.0 of an ability parameter estimate is 40.0 and the ninety percent confidence interval is between 34.9 and 45.1. When the ability parameter estimate is 0.0, the Z-score is 50.0 and the confidence interval is between 44.6 and 55.4, the Z-score at 1.0 of an ability parameter estimate is 60.0, of which the confidence interval is between 53.1 and 66.7. Since the degree of confidence is 90%, which is high enough, the three examples noted above may give the impressions that the confidence interval is relatively wide. As the number of items administered to test takers increases, the standard error of the ML estimate decreases. In order to reduce the range of confidence intervals with a 90% degree of confidence, meaning that the precision of ability parameter estimates have been raised, more than 50 items will be needed in one test.
3.3 Computer Program

An executable computer program, EST3.EXE, was developed which computes the ability parameter estimates for the one-parameter logistic, the two-parameter logistic, and the three-parameter logistic item response theory models. EST3.EXE calculates the item response theory maximum-likelihood and Bayesian estimates for a test administered by the paper-and-pencil method. The output from EST3.EXE contains maximum-likelihood, Bayesian-modal, and EAP ability estimates for each test taker. EST3.EXE requires discriminating (a), difficulty (b), and guessing(c) item parameters and test taker response data as input. The mean and variance of a prior distribution of the ability parameter can be entered via keyboard. An example of a command file with item response data and output from the EST3.EXE program saved in an external file is illustrated in Appendix 2.

4. Discussion

The item bank developed in this study can be a useful and convenient tool for schoolteachers to create English tests at either a class-level or a large-scale level. There are a number of advantages to using the item bank for creating tests.

First, it makes it easier for teachers to create tests. Teachers can construct different versions of tests from the item bank according to the purpose of test. Tests with different levels of difficulty or parallel tests are easily created using the item parameters.

Second, the ability of the students can be estimated on a common scale and become comparable even though each student may take a different test with different combinations of items drawn from the item bank. Therefore, there are many situations at schools where teachers can make use of the item bank. Examples include: (1) learning the effects of different teaching methods on different classes; (2) discovering the change in students’ English ability from one year to the next; (3) tracking the change of any individual student’s English ability throughout his/her three high school years; (4) comparing the ability of students who entered high school through the regular entrance examination with that of students who entered by means of a recommendation; (5) comparing the ability of a student who wants to transfer to a certain school with that of students already attending the school.

Third, the item bank can be expanded when administering a large-scale test consisting of both new items and items from the item bank. The item parameters of new items can be estimated using the item parameters in the item bank, and then new items can be added into the item bank. In this way, the item bank can be updated.

This study provides high school teachers with an item bank which can be used to create classroom English tests or develop large-scale English tests for practical use. It also offers a user-friendly computer program to estimate ability parameters from item bank testing. For these reasons, we believe that this study is significant.
The following should be considered for further study. First, from a technical perspective, making the item bank more accessible needs to be considered. Second, it will be necessary to expand the item bank, adding a variety of items to measure other abilities such as communicative competence more precisely. Third, more than 20 prefectures have administered prefecture-wide English tests like Ibaraki’s (Saida, 2002). If these prefecture-wide tests contain some items from this item bank, the results will be comparable on a common scale, which will enable the establishment of a huge educational data bank that can be applied beyond prefectural borders.

Notes
(1) Item bank
Items and item parameter estimates are available from Chisato Saida at Namiki High School, Ibaraki: 4-5-1 Namiki, Tsukuba, Ibaraki, 305-0044. E-mail:saida@namiki-h.ed.jp
(2) Computer program
The executable EST3.EXE program, Japanese user’s guide in PDF format, and a set of example input and output files are available without cost from Tamaki Hattori, University of Tsukuba, Institute of Psychology, Tsukuba Ten-noudai 1-1-1, Ibaraki 305-8572, Japan, E-mail:hattori@human.tsukuba.ac.jp. It also can be downloaded from the following URL. http://www.human.tsukuba.ac.jp/~hattori/irt/irt.html

References
大友賢二. (1996). 「項目応答理論入門—言語テスト・データの新しい分析法—」
Appendix 1. Examples of items and item parameters

Listening

2000(3) \( (a=0.60 \ b=1.41) \)

A: Happy birthday, Bob. This is for you! B: For me?
A: Today’s April 16th and it’s your birthday, isn’t it? B: No, Cindy. It was three days ago.
Q: When is Bob’s birthday?

1. April 3rd. 2. April 13th. 3. April 16th. 4. April 19th.

Vocabulary

2001(19) \( (a=0.85 \ b=0.24) \)

May I ( ) this telephone?
1. speak. 2. call. 3. use. 4. talk

Dialogue

1999(23) \( (a=0.88 \ b=1.48) \)

"Why don’t you come to my house?"
1. I didn’t go to your house. 2. Because I’m late.
3. Yes, I’d love to. 4. Because I was busy.

Grammar

1997(30) \( (a=1.22 \ b=-0.55) \)

She doesn’t know ( ).
1. where I live. 2. I live how.
3. where do I live. 4. how do I live
Appendix 2. Computer Program EST3.EXE

(1) Command file formatted in ASCII file

<table>
<thead>
<tr>
<th>Item</th>
<th>A_j</th>
<th>B_j</th>
<th>C_j</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.55</td>
<td>1.49</td>
<td>1.42</td>
</tr>
<tr>
<td>2</td>
<td>1.41</td>
<td>-1.56</td>
<td>1.88</td>
</tr>
<tr>
<td>3</td>
<td>1.11</td>
<td>-0.88</td>
<td>1.48</td>
</tr>
<tr>
<td>4</td>
<td>1.07</td>
<td>-1.48</td>
<td>0.92</td>
</tr>
<tr>
<td>5</td>
<td>0.73</td>
<td>0.92</td>
<td>1.50</td>
</tr>
<tr>
<td>6</td>
<td>0.47</td>
<td>1.43</td>
<td>1.20</td>
</tr>
<tr>
<td>7</td>
<td>1.46</td>
<td>1.32</td>
<td>1.18</td>
</tr>
<tr>
<td>8</td>
<td>1.50</td>
<td>1.18</td>
<td>0.78</td>
</tr>
<tr>
<td>9</td>
<td>1.91</td>
<td>0.78</td>
<td>0.00</td>
</tr>
</tbody>
</table>

This test consists of five items which were calibrated by DPL. There are discriminating and difficulty values from left to right in each row.

"Y" indicates the column which contains item response and "I" identification character. Correct key vector
Not-presented key vector
Response vector and identification characters

(2) Output from the EST3.EXE program

EST3:
Program for estimating ability parameters of logistic models (D is fixed at 1.0)
by T. Hattori : 8/3/2002

Number of item parameters = 2

Item parameters

<table>
<thead>
<tr>
<th>Item</th>
<th>A_j</th>
<th>B_j</th>
<th>C_j</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.55</td>
<td>1.49</td>
<td>1.42</td>
</tr>
<tr>
<td>2</td>
<td>1.41</td>
<td>-1.56</td>
<td>1.88</td>
</tr>
<tr>
<td>3</td>
<td>1.11</td>
<td>-0.88</td>
<td>1.48</td>
</tr>
<tr>
<td>4</td>
<td>1.07</td>
<td>-1.48</td>
<td>0.92</td>
</tr>
<tr>
<td>5</td>
<td>0.73</td>
<td>0.92</td>
<td>1.50</td>
</tr>
<tr>
<td>6</td>
<td>0.47</td>
<td>1.43</td>
<td>1.20</td>
</tr>
<tr>
<td>7</td>
<td>1.46</td>
<td>1.32</td>
<td>1.18</td>
</tr>
<tr>
<td>8</td>
<td>1.50</td>
<td>1.18</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Specification list for items to be included and identification fields for subjects:
YYYYYYYYYY IIIIIIIII

Correct answer key vector:
3423423423

Not-presented key vector:
9999999999

Mean of prior distribution of theta = 0.000
Variance of prior distribution of theta = 1.000

Maximum likelihood and Bayes estimates
Resp. is responses of the first five items.
Tried is the number of tried items.
Right is the number of correct responses.
MLE is the maximum likelihood estimate.
MODAL is the Bayes modal estimate.
EAP is the expected a posteriori estimate.
SE is the standard error of theta.

+/-99.999 indicates that the estimate was not computed.
+/-88.999 indicates that the solution diverged.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Resp.</th>
<th>Tried</th>
<th>Right</th>
<th>MLE</th>
<th>SE</th>
<th>MODAL</th>
<th>SE</th>
<th>EAP</th>
<th>SE</th>
<th>Identification characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3423</td>
<td>8</td>
<td>6</td>
<td>0.97</td>
<td>0.79</td>
<td>0.593</td>
<td>0.628</td>
<td>0.587</td>
<td>0.635</td>
<td>2-1 1</td>
</tr>
<tr>
<td>2</td>
<td>1213</td>
<td>10</td>
<td>5</td>
<td>-0.585</td>
<td>0.695</td>
<td>-0.392</td>
<td>0.578</td>
<td>-0.375</td>
<td>0.580</td>
<td>2-1 2</td>
</tr>
<tr>
<td>3</td>
<td>2123</td>
<td>10</td>
<td>6</td>
<td>-0.551</td>
<td>0.697</td>
<td>-0.368</td>
<td>0.579</td>
<td>-0.351</td>
<td>0.580</td>
<td>2-1 3</td>
</tr>
<tr>
<td>4</td>
<td>4912</td>
<td>8</td>
<td>4</td>
<td>-0.311</td>
<td>0.768</td>
<td>-0.195</td>
<td>0.613</td>
<td>-0.182</td>
<td>0.609</td>
<td>2-1 4</td>
</tr>
<tr>
<td>5</td>
<td>2223</td>
<td>7</td>
<td>5</td>
<td>0.086</td>
<td>0.991</td>
<td>0.0448</td>
<td>0.663</td>
<td>0.087</td>
<td>0.659</td>
<td>2-1 5</td>
</tr>
</tbody>
</table>