Practical use of computerized adaptive testing in Japan

N. Hayashi, Y. Nogami, K. Maeda & H. Ikeda

(The Japan Institute for Educational Measurement, Inc.)

ICP 2004
Beijing, China

August 13th, 2004
• Introduction of CASEC as an example of Computerized Adaptive Testing (CAT) in Japan.
  – Demonstration
  – Item Bank Development
  – Ability Estimation Method & Item Selection Rule

• Two experiments
  – CAT is more efficient than Paper-and-Pencil (P&P) test.
  – the reliability of CASEC.

• Summary
What is CASEC?

- Computerized Assessment System for English Communication (CASEC).
- Developed at JIEM.
- A test system to measure English proficiency.
- CAT based on Item Response Theory.
- Employed in Japan, mainly at universities, companies, and junior high schools.
- Web version & CD-ROM version.
- The Japan Institute for Educational Measurement, Inc.
Section 1
- [Image]

Section 2
- [Image]

Section 3
- [Image]

Section 4
- [Image]
CASEC Score Report

<table>
<thead>
<tr>
<th>受験日</th>
<th>番号</th>
<th>注文番号</th>
<th>Sec1</th>
<th>Sec2</th>
<th>Sec3</th>
<th>Sec4</th>
<th>Total</th>
<th>TOEIC</th>
<th>TOEFL(%)</th>
<th>改検級</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001/07/26</td>
<td>211</td>
<td>204</td>
<td>229</td>
<td>174</td>
<td>818</td>
<td>300</td>
<td>551</td>
<td>出力部</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

アドバイス
現在の語彙や表現の状況を活かすことで、円滑なコミュニケーションが可能で、よりリスニング力も上上がるので、自信を持って積極的にコミュニケーションを心がけましょう。また、英語のある分野から始めて、より専門的な語彙や表現を学習すると、話題の幅が広がり、深いのある会話ができるようになります。

CASECスコア（見出し）

- CASECスコアとは、英語のコミュニケーション能力を評価するテストです。
- スコアは、Sec1-4の合計スコアとTotalスコアが示されます。
- 各Secのスコアは、最低0から最高30までのスケールで評価されます。
- Totalスコアは、Sec1-4の合計スコアで、最低0から最高120までのスケールで評価されます。

報告の前回と今回を比較

- 前回の試験結果と今回の中間の試験結果を比較して、進歩を確認できます。
- 比較の対象は、Sec1-4のスコアとTotalスコアです。
Item Bank of CASEC

- 4,000 items
- 3-Parameter Logistic Model (Sections 1-3)
- 2-Parameter Logistic Model (Section 4)

To obtain those parameters...

- more than 10 pre-tests with Common Item Design
- 1 to 6 forms (120 items/form) for each pre-test
- at least 1,000 examinees for each test form

Equating method (mean-sigma method) (Marco, 1977; Cook & Eignor, 1991)
Ability Estimation and Item Selection Rule

- Item Selection Rule
  - Which item should be presented as the next item?

\[ \hat{\theta}_{t-1} - 0.5 < b_t < \hat{\theta}_{t-1} + 0.5 \]

- Ability Estimation Method:
  - EAP (expected a posteriori; Bock & Mislevy, 1982)

\[ \hat{\theta}_{t-1} - 0.5(\text{SE}_{\hat{\theta}})_{t-1} < b_t < \hat{\theta}_{t-1} + 0.5(\text{SE}_{\hat{\theta}})_{t-1} \]
• the Initial value of $\theta$ (ability)

\[ \hat{\theta}_{t-1} - 0.5 < b_t < \hat{\theta}_{t-1} + 0.5 \]

• The difficulty parameter of the first item

\[ -0.5 < b_1 < 0.5 \]

\[ \hat{\theta}_0 - 0.5 < b_1 < \hat{\theta}_0 + 0.5 \]

\[ \hat{\theta}_0 = 0 \]
Experiments

• We did many experiments while developing CASEC.
• 2 of them will be reported.

• Purpose of Exp1
  – Examine advantages (precision, etc...) of CASEC compared to P&P.

• Purpose of Exp2
  – Examine reliability of CASEC.
Testing tools

• infinite

was defined based on

a Stopping Rule – \( SE \text{Į} < 0.05 \) & \( SE \text{Įt} - SE \text{Įt-1} < 0.001 \)

– 30 items (One Section)

• 1 form

120 items

(30 items / One Section)

Item Bank

CAT only

1 form

P&P

Test Length

Number of Forms

Same Scoring Method

Scores from both tests were comparable
168 examinees took both CASEC and P&P in the same day.

The orders were counter-balanced.

N = 168
- **Test Length**: CAT < P&P
- **Standard Error**: CAT < P&P

**Advantage of CAT Test Length**

**Mean Test Lengths of P&P and CAT**

**Mean SEs of P&P and CAT**
• 48 examinees took CASEC 3-times in the day.
Exp2: Reliability of CASEC

Although each test consists of different items, a degree of coincidence was observed for the scores between each trial.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>423.2</td>
<td>2.6</td>
</tr>
<tr>
<td>2</td>
<td>425.2</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>423.7</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Correlations between trials (CAT)

Means and SDs of total scores

(N=48)

- CASEC has very high reliability!!
Improving CASEC

• Adding more items to the item bank
  – items are not enough to use more complicated item selection rules

• Methods to set the initial value of theta
  – Using available information is important
The initial value is far from the examinee's ability.

**Typical Case**

- **Slow Convergence Rate Case**
- **Rapid Convergence**
- **Fast Convergence Rate Case**
• CASEC was introduced as an example of CAT in Japan
• The results of the experiments conducted on CASEC were presented.
• We have demonstrated that CASEC is a very effective technique in terms of measurement accuracy and reliability.

