UDC 519.6

テストレットモデルに基づく適応型テストシステム

石塚　智一*
前川　眞一**
菊地　賢一***
中畝菜穂子*
内田　照久****

要　約

テストレットモデルに基づく適応型のCBT (Computer Based Testing)システムを構築し，その性能を2つの方法で評価した．第1は実受験者の解答パタンに基づくもの，第2は正規乱数によって発生した受験者の能力値に基づくものである．第1，第2の方法を通じて真の能力値と推定された能力値の間に高い相関が見られたので，システムはほぼ期待通りに働くことが確認された．

1　はじめに[1]

被験者の反応によって次に出題される問題が異なる，枝分かれテスティング(Branched Testing)，プログラム・テスティング(Programmed Testing)の試みは，1960年代にも行われているが(Linn, Rock, and Cleary; 1969)，これを項目反応理論(Item Response Theory)と結びつけたのは恐らくLord (1970)が初めであろう．被験者が答えた項目に依存せずにその能力の推定を可能とする項目反応理論は，その後適応型テスト(Adaptive Testing)システムとは不可分のものとなる．

一方，Wainer and Kiely (1987)は，CBTにおける適応形テストの文脈から，幾つかの項目を一塊の項目群として捉えることの利点を論じ，これをテストレット(Testlet)と呼んだ．彼等は，テストレットのタイプとして，階層型テストレット(Hierarchical Testlet)，線型テストレット(Linear Testlet)，及びその混合型について論じた．

大問形式の問題や長文読解のように1つの長文に幾つかの設問が付属する形式の問題では，設問間の共分散に，受験者の能力を反映した共分散の他に，大問の出題範囲や長文のトピックなどに由来する共分散が混じり，設問間に局所独立性を仮定することができない．このようなテストに局所独立性を仮定する項目反応理論のモデルを当てはめると能力推定の標準誤差が過小に評価される(Thissen, Steinberg, and Moony; 1989)．

テストレットの考え方では，テストレット内の設問間に局所独立性を仮定しなくても，テストレットの間に局所独立性が仮定できれば項目パラメタ

* 大学入試センター研究開発部 試験作成支援研究部門
〒153–8501 東京都目黒区駒場 2–29–23　ishizuka@rd.dnc.ac.jp
** 東京工業大学大学院 社会理工学研究科
*** 東邦大学 理学部
**** 大学入試センター研究開発部 試験環境研究部門
2002年8月6日　受理
[1]本研究は文部科学省科学研究費(基盤研究C「CAT開発へ向けての項目プールの作成—大問反応理論による試験問題の分析」研究代表者：石塚智一 課題番号 12480047)の補助により実施された．

や受験者の能力の推定が可能となる(Wainer and Kiely, 1987). Rosenbaum (1988)はテストレットが局所独立になる条件について議論している.

我が国の試験は大問形式で出題されることが多い. そこで, その大問を1つのテストレットと考えれば, これは彼等のいう線型テストレットに当たるものと考えられる. 石塚他(2001)はこれを我が国の英語の試験問題に当てはめ, その性質を評価した.

石塚他(2001)は英語の出題の内, 設問間に局所独立性を仮定できない大問については, それをテストレットと考え, その大問への正答数を以ってテストレットの得点とした. こうして, 各テストレットは, そこに含まれる設問数+1の得点段階を持つことになる. このようなテストレットの反応モデルとしてはGraded Response Model (Samejima, 1969, 1972)が相応しいと考えた.

我が国の試験作成が大問形式に依存する限り, 適応型CBTシステムもこれに対応したものを用意する必要がある. 石塚他(2002) はテストレットモデルに基づく適応型CBTシステムの構築を試み, その評価を行っている. 本稿は石塚他(2002)を基に幾つかの改善を試みたものである.

2 準 備

アルゴリズムの記述に立ち入る前に, 必要となるテストレット反応モデル(Testlet Response Model)の概念を幾つか定義しておく.

テストレット応答関数(Testlet Response Function)

Graded Response Modelのテストレット応答関数は以下の式で与えられる. ここで, $P_{jk}(\theta)$ は特定の θ を持つ受験者が j 番目のテストレットで k 点を獲得する確率を示す. すなわち, j 番目のテストレットに含まれる項目数を m_j で表して

$$P_{jk}(\theta) = P^{+}_{jk}(\theta) - P^{+}_{j,k+1}(\theta), \quad k = 0, 1, \ldots, m_j$$

ここで

$$P^{+}_{j0}(\theta) = 1$$
$$P^{+}_{j,m_j+1}(\theta) = 0$$

および

$$P^{+}_{jk}(\theta) = \frac{1}{1 + \exp[-Da_j(\theta - b_{jk})]}$$

ここで, a_j は項目識別力に相当するパラメタ, b_{jk} は j 番目のテストレットにおける k 番目のカテゴリの困難度を定めるパラメタである. また, D はロジスティック関数を累積正規曲線に近似させるための定数で $D = 1.7$ である.

テストレット情報関数(Testlet Information Function)

テストレット情報関数は受験者の能力値 θ におけるテストレットの測定精度を表すもので, 以下のように与えられる.

$$I_j(\theta) = \sum_{k=0}^{m_j} A_{jk}(\theta)$$

ただし

$$A_{jk}(\theta) = D^2 a_j^2 \cdot \frac{\{P^{+}_{jk}(\theta)[1 - P^{+}_{jk}(\theta)] - P^{+}_{j,k+1}(\theta)[1 - P^{+}_{j,k+1}(\theta)]\}^2}{P_{jk}(\theta)}$$

反応パタンの尤度(Likelihood of Response Pattern)

t 回目の試行までの反応パタン V_t の尤度は

$$L[V_t \mid \theta] = \Pi_{j=1}^{t} \Pi_{k=0}^{m_j} [P_{jk}(\theta)]^{U_{jk}}$$

で与えられる. U_{jk} は j 番目のテストレットにおいて, 得点が k であれば1, それ以外であれば0を取る関数である.

能力の事後分布の期待値(EAP: Expected a Posteriori)

t 試行後の能力の事後分布の期待値は

$$\bar{\theta}_t = \frac{\sum_{l=1}^{q} X_l L(X_l) W(X_l)}{\sum_{l=1}^{q} L(X_l) W(X_l)}$$

で与えられる. これは, θ の事前分布を $w(\theta)$ として得られる事後分布の期待値

$$\bar{\theta}_t = \frac{1}{\int_{-\infty}^{\infty} L(V_t \mid \theta) w(\theta) d\theta} \times \int_{-\infty}^{\infty} \theta L(V_t \mid \theta) w(\theta) d\theta$$

の積分を数値積分で置き換えたものである．また，X_l は数値積分のための分点で，$W(X_l)$ は X_l における重みである．ここでは，-3 から $+3$ までの間に 61 の分点を設け，重みは標準正規分布における各分点の密度を，その和が 1 となるように規準化して用いている．

事後分布の標準偏差(PSD: Posterior Standard Deviation)

$$PSD(\theta)_t = \left[\frac{\sum_{l=1}^{q} (X_l - \bar{\theta}_t)^2 L(X_l) W(X_l)}{\sum_{l=1}^{q} L(X_l) W(X_l)} \right]^{\frac{1}{2}}$$

これも事後分散

$$PV = \frac{1}{\int_{-\infty}^{\infty} L(V_t \mid \theta) w(\theta) d\theta} \times \int_{-\infty}^{\infty} (\theta - \bar{\theta}_t)^2 L(V_t \mid \theta) w(\theta) d\theta$$

の積分を数値積分で置き換え，平方根をとったものである．

3 システムの概要

3.1 項目プール

項目プールは平成 12 年度に実施された大学入試センター試験の英語の出題をテストレット化したもの(石塚他，2001)を利用する．この理由は，すでに実受験者の解答パタンや能力の推定値が存在しているので，システムの評価に当たって都合が良いと考えられるからである．

一方，このデータは 29 個のみのテストレットからなるので，項目プールとしてはあまりに小さく，別の観点からは不都合がある．そこで，400 個の大きさと 800 個の大きさの項目プールを一定の基準に基づいて確率的に生成した．このための基準や，項目プール生成のアルゴリズムについては後述する．

3.2 テストレットの選択

受験者の能力の推定値においてその能力値にとって最適なテストレットを選択して提示できることが適応型テストのセールスポイントである．テストレットの選択方略としては，PSD の期待値を最小とするようなテストレットを選ぶのが能力のベイズ的推定法と整合的だが，PSD は受験者の解答が与えられてはじめて得られるものである．受験者の解答と能力推定値が得られる度に項目プールを検索するのは効率が悪い．PSD の期待値を最小にするテストレットを選ぶことと，受験者の能力値における情報量を最大にするテストレットを選ぶことは，ほぼ同じ意味をもっている．そこで，試験の開始前に値が得られる情報量を利用して以下のような便法を用意する．まず，能力尺度の -4 から $+4$ の間を 81 の能力段階に量子化する．次に，各能力段階毎にプールに存在する全てのテストレットの情報量を計算し，予め情報量の大きい順に並べ替えた表を作っておく．これを情報量表と呼ぶ(図 1)．$\bar{\theta}_t$ が得られたならば，情報量表において $\bar{\theta}_t$ に最も近い能力段階に当たる行から，未使用テストレットを情報量の大きい順に選んで提示する．

3.3 ストッピングルール

各受験者の能力推定は，PSD が 0.3162 を下回るか，項目プールの底をついた時に終了する．PSD の 0.3162 とは，そこでの情報量が 10 となる値で，信頼性にして 0.9 に当たる精度である(Thissen and Mislevy, 2000)．

3.4 アルゴリズム

0. $\bar{\theta}_1 = 0$ とする．
1. 情報量表を用いてにおいて $\bar{\theta}_t$ 情報量が最大となる未使用テストレットを選択する．
2. 受験者の解答を得る(実験 I では実受験者の解答を，実験 II ではテストレット応答モデルに基づいて確率的に生成した解答を利用する)．
3. $\bar{\theta}_{t+1}$ および PSD を計算する．
4. PSD が 0.3162 より小さいか項目プールの終わりとなればその受験者を終了し，次の受験者にすすむ．そうでなければ，$t = t + 1$ として 2. へ戻る．
5. 以上を受験者の数だけ繰り返す．

−4.0	29	28	27	20	19	8	4	2	3	26	13	12	15	11	5	24	17	1	21	6	9	25	10	18	14	7	23	22	16
−3.9	29	28	27	20	19	4	8	2	3	26	13	12	11	15	24	17	5	21	1	6	9	25	10	18	14	7	22	23	16
−3.8	29	28	27	20	19	4	2	8	3	13	26	12	11	24	15	17	21	5	6	1	9	25	10	18	14	7	22	23	16
−3.7	29	28	27	20	19	4	2	3	13	8	26	11	12	24	15	21	17	5	6	9	1	25	10	18	14	7	22	23	16
−3.6	29	28	27	20	19	4	13	2	3	26	8	11	24	12	15	21	17	6	5	9	25	1	10	18	14	7	22	23	16
−3.5	29	28	27	20	19	13	4	2	3	26	11	24	8	12	15	21	17	6	5	9	25	1	10	18	14	7	22	23	16
−3.4	29	28	27	20	19	13	4	3	2	11	24	26	12	8	21	15	17	6	9	25	5	1	10	18	14	7	22	23	16
−3.3	29	28	27	20	19	13	11	24	4	3	2	26	12	21	15	8	17	6	9	25	5	1	10	18	14	22	7	23	16
−3.2	29	28	27	20	19	13	11	24	3	4	26	2	21	12	15	8	6	17	9	25	5	1	10	18	14	22	7	23	16
−3.1	29	28	27	20	19	13	11	24	26	3	4	2	21	12	15	6	8	17	25	9	5	1	10	18	14	22	7	23	16
−3.0	29	28	27	20	19	13	11	24	26	3	2	4	21	12	15	6	17	8	25	9	5	1	10	18	14	22	7	23	16
−2.9	29	28	27	20	19	13	24	11	26	3	2	4	21	12	15	6	17	25	9	8	5	10	1	18	22	14	7	23	16
−2.8	29	28	27	20	13	24	11	19	26	3	21	2	4	12	15	6	17	25	9	8	5	10	1	22	18	14	7	23	16
−2.7	29	28	27	24	20	11	13	19	26	21	3	2	4	12	15	6	17	25	9	8	5	10	22	1	18	14	7	23	16
−2.6	29	28	27	24	11	13	20	19	21	26	3	2	4	12	15	6	25	17	9	8	22	5	10	1	18	14	7	23	16
−2.5	29	28	27	24	11	13	20	19	21	26	3	2	4	12	15	6	25	17	9	22	8	5	10	1	18	14	7	23	16
−2.4	29	28	27	24	11	13	20	19	21	26	3	2	4	12	15	6	25	17	9	22	8	5	10	1	18	14	7	23	16
−2.3	29	28	27	24	11	13	20	21	19	26	3	2	4	12	6	15	25	17	9	22	8	10	5	1	18	14	7	23	16
−2.2	29	28	27	24	11	13	20	21	19	26	3	2	4	12	6	15	25	9	17	22	10	5	8	18	1	14	7	23	16
−2.1	29	28	27	24	11	13	21	20	19	26	3	2	4	6	12	15	25	22	9	17	10	5	8	18	1	14	7	23	16
−2.0	29	28	27	24	11	13	21	20	19	26	3	2	6	12	4	15	22	25	9	17	10	5	18	8	1	14	7	23	16
−1.9	29	28	24	27	11	13	21	20	26	19	3	2	22	6	12	25	15	4	9	17	10	5	18	8	1	14	7	23	16
−1.8	29	28	24	27	11	13	21	20	26	19	3	22	6	2	25	12	15	4	9	17	10	5	18	1	8	14	7	23	16
−1.7	29	28	24	27	11	13	21	26	20	22	3	19	6	2	25	12	15	4	9	17	10	5	18	14	1	8	7	23	16
−1.6	29	28	24	27	11	13	21	22	26	20	3	19	6	25	2	12	15	4	9	17	10	5	18	14	1	8	7	23	16
−1.5	29	28	24	27	11	13	21	22	26	3	20	6	19	25	2	15	12	9	4	17	10	5	18	14	1	7	8	23	16
−1.4	29	28	24	27	11	13	21	22	26	3	6	20	25	19	2	15	12	9	4	17	10	5	18	14	1	7	8	23	16
−1.3	29	28	24	27	11	21	13	22	26	3	6	25	20	19	15	2	12	9	4	17	10	18	5	14	7	1	8	23	16
−1.2	29	28	24	27	11	22	21	13	26	3	6	25	20	15	19	9	12	2	17	4	10	18	5	14	7	1	8	23	16
−1.1	29	28	24	27	22	11	21	13	3	26	6	25	9	15	20	12	2	19	17	4	10	18	5	14	7	1	8	23	16
−1.0	29	28	24	27	22	11	21	13	3	6	25	26	9	15	12	20	2	19	17	4	10	18	5	14	7	1	8	23	16
−.9	29	28	24	27	22	11	21	13	3	25	6	26	9	15	12	2	20	17	19	4	10	18	5	14	7	1	8	23	16
−.8	29	28	24	22	27	11	21	13	3	25	6	26	9	15	12	2	17	20	19	4	10	18	5	14	7	1	8	23	16
−.7	29	28	24	22	27	21	11	13	3	25	6	26	9	15	12	2	17	20	19	4	10	18	5	14	7	1	8	23	16
−.6	29	28	24	22	27	21	11	13	3	25	6	26	9	15	12	17	2	20	19	4	10	18	5	7	14	1	8	23	16
−.5	29	28	22	24	27	21	11	13	3	25	6	26	9	15	12	17	2	20	19	4	10	18	5	7	14	1	8	23	16
−.4	29	28	22	24	27	21	11	3	25	13	6	9	26	15	12	17	2	4	19	20	10	18	7	5	14	1	8	23	16
−.3	29	28	22	24	27	21	11	3	25	6	13	9	26	15	12	17	2	4	19	20	10	18	7	5	14	1	8	23	16
−.2	29	28	22	24	27	21	3	11	25	6	13	9	26	15	17	12	2	4	10	19	20	7	18	5	14	1	8	23	16
−.1	29	28	22	24	27	21	3	25	6	11	9	13	26	15	17	12	2	10	4	7	18	19	5	20	14	1	8	23	16
.0	29	28	22	24	27	3	21	25	6	11	9	13	26	15	17	12	2	10	7	4	18	5	14	19	20	1	8	23	16
.1	29	28	27	24	22	3	21	25	6	9	11	13	15	26	17	12	2	7	10	18	5	4	14	19	20	1	8	23	16
.2	29	28	27	24	22	3	21	25	6	9	11	15	13	26	17	12	7	2	10	18	5	14	4	19	20	1	23	8	16
.3	29	28	27	24	22	3	25	21	6	9	11	15	17	26	13	12	7	10	2	18	5	14	4	19	20	1	23	8	16
.4	29	28	27	24	22	3	25	21	6	9	15	11	17	26	13	12	7	10	18	5	2	14	4	19	20	1	23	8	16
.5	29	28	27	24	22	3	25	6	21	9	15	17	7	11	26	12	13	10	5	18	14	2	4	19	20	1	23	8	16
.6	29	28	27	24	22	3	25	6	9	21	15	17	7	26	12	11	10	5	18	13	14	2	4	19	1	20	23	8	16
.7	29	28	27	24	22	3	25	6	9	21	7	15	17	12	26	10	5	18	14	11	13	2	4	1	19	20	23	8	16
.8	29	28	27	24	22	3	25	6	9	21	7	15	17	5	10	12	18	14	26	11	13	2	4	1	19	20	23	8	16
.9	29	28	27	24	22	3	25	6	9	7	21	17	15	5	10	18	14	12	26	11	2	13	4	1	19	20	23	8	16
1.0	29	27	28	24	3	22	25	9	6	7	21	17	15	5	10	18	14	12	26	2	11	13	4	1	19	20	23	8	16
1.1	29	27	28	24	3	22	25	9	6	7	5	21	17	15	10	14	18	12	26	2	13	11	4	1	19	20	23	8	16
1.2	29	27	28	24	3	22	25	9	7	6	5	17	14	10	18	15	21	12	26	2	13	11	4	1	19	23	20	8	16
1.3	29	27	28	3	24	22	25	7	9	6	5	14	17	18	10	15	21	12	26	2	4	13	1	11	23	19	20	8	16
1.4	29	27	28	3	24	22	7	25	9	6	5	14	18	10	17	15	21	12	26	2	4	1	13	11	23	19	20	8	16
1.5	29	27	28	3	24	22	7	25	9	5	6	14	18	10	17	15	21	12	26	2	1	4	13	11	23	19	20	8	16
1.6	29	27	28	3	24	22	7	25	9	5	6	14	18	10	17	15	12	21	26	2	1	4	23	13	11	19	20	8	16
1.7	29	27	28	3	24	22	7	25	5	9	6	14	18	10	17	15	12	21	26	2	1	4	23	13	11	19	20	8	16
1.8	29	27	28	3	24	7	22	5	25	9	14	6	18	10	17	15	12	21	26	2	1	23	4	13	11	19	20	8	16
1.9	29	27	28	3	24	7	5	22	25	9	14	18	6	10	17	15	12	21	26	1	2	23	4	13	11	19	20	8	16
2.0	29	27	28	3	7	24	5	25	22	9	14	18	10	6	17	15	12	21	26	1	23	2	4	13	11	19	20	8	16
2.1	29	27	28	3	7	24	5	14	25	9	18	22	10	6	17	15	12	23	1	21	26	2	4	13	11	19	20	8	16
2.2	29	27	3	28	7	5	24	14	25	18	9	10	22	6	17	15	12	23	1	21	26	2	4	13	11	19	20	8	16
2.3	29	27	3	28	7	5	24	14	18	9	25	10	6	22	17	15	12	23	1	26	21	2	4	13	19	11	20	8	16
2.4	29	27	3	28	7	5	14	24	18	10	9	25	6	17	22	15	23	12	1	26	2	21	4	13	19	11	20	8	16
2.5	29	27	3	28	7	5	14	18	24	10	9	25	6	17	22	15	23	12	1	26	2	21	4	13	19	11	20	8	16
2.6	27	29	3	28	7	5	14	18	10	9	24	25	6	17	15	22	23	12	1	26	2	21	4	13	19	11	20	8	16
2.7	27	29	3	28	7	5	14	18	10	9	25	24	17	6	23	15	22	12	1	2	26	4	21	13	19	11	20	8	16
2.8	27	29	3	7	28	5	14	18	10	9	25	24	17	6	23	15	22	1	12	2	26	4	21	13	19	8	11	20	16
2.9	27	29	3	7	5	28	14	18	10	9	25	24	17	23	6	15	1	12	22	2	26	4	21	19	13	8	20	11	16
3.0	27	29	3	7	5	28	14	18	10	9	25	23	17	24	6	15	1	12	22	2	26	4	21	19	13	8	20	11	16
3.1	27	29	3	7	5	28	14	18	10	9	25	23	17	6	24	15	1	12	22	2	26	4	21	19	13	8	20	11	16
3.2	27	29	3	7	5	14	28	18	10	9	23	25	17	6	24	15	1	12	22	2	26	4	21	19	8	13	16	20	11
3.3	27	29	3	7	5	14	18	28	10	23	9	25	17	6	1	15	24	12	22	2	26	4	21	19	8	13	16	20	11
3.4	27	29	3	7	5	14	18	28	10	23	9	25	17	6	1	15	24	12	22	2	26	4	21	19	8	16	13	20	11
3.5	27	29	3	7	5	14	18	10	28	23	9	25	17	6	1	15	24	12	2	22	4	26	21	16	8	19	13	20	11
3.6	27	29	3	7	5	14	18	10	23	28	9	25	17	1	6	15	12	24	2	4	26	22	21	16	8	19	13	20	11
3.7	27	29	3	7	5	14	18	10	23	28	9	25	17	1	6	15	12	24	2	4	26	22	21	16	8	19	13	20	11
3.8	27	29	3	7	5	14	18	23	10	28	9	17	25	1	6	15	12	24	2	4	26	22	21	16	8	19	13	20	11
3.9	27	29	3	5	7	14	18	23	10	9	28	17	25	1	6	15	12	24	2	4	26	22	21	16	8	19	13	20	11
4.0	27	5	7	3	29	14	18	23	10	9	28	17	25	1	6	15	12	24	2	4	26	22	16	21	8	19	13	20	11

図 1　情報量表の例

表 1 項目プールのパラメタ(実験 1)*

項目	大きさ	a	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
T1	2	0.288	−2.517							
T2	2	0.523	−2.086							
T3	3	0.549	−1.963	0.930						
T4	2	0.498	−2.368							
T5	4	0.293	−9.080	−2.401	1.723					
T6	2	0.556	−1.053							
T7	2	0.348	1.090							
T8	2	0.407	−4.452							
T9	2	0.494	−0.888							
T10	2	0.344	−0.804							
T11	2	0.866	−1.645							
T12	2	0.497	−1.726							
T13	2	0.780	−1.782							
T14	2	0.309	−0.061							
T15	2	0.497	−1.466							
T16	2	0.063	14.858							
T17	2	0.449	−1.335							
T18	2	0.323	−0.473							
T19	2	0.620	−2.550							
T20	2	0.662	−2.561							
T21	2	0.751	−1.318							
T22	2	0.937	−0.387							
T23	2	0.170	2.915							
T24	3	0.880	−1.637	−0.184						
T25	2	0.555	−0.822							
T26	2	0.594	−1.776							
T27	6	0.798	−2.491	−1.152	−0.103	0.917	2.160			
T28	6	0.931	−3.568	−2.428	−1.494	−0.599	0.501			
T29	9	1.056	−3.774	−2.656	−1.822	−1.155	−0.575	0.013	0.708	1.625

* 表 1 の内容は石塚他(2001)より転載

4 実受験者の解答を用いたシステムの評価(実験 I)

4.1 受験者

平成 12 年度に英語を受験した者から無作為に 1002 名を選んだ．外国語の成績ファイルからサンプリングを行ったため，英語以外の科目の受験生がサンプルに混じることを心配して 2 名分余計にサンプリングしたが，結局選ばれたのが全て英語の受験者であったので 2 名の端数が生じた．

4.2 項目プール

この年度の出題をテストレット化したものを項目プールとする．実際の出題とテストレットの対応は石塚他(2001)に詳しく掲載されているが，プールの中のテストレットのパラメタを表 1 に示す．また，このプールから作成した情報量表を表 2 に示す．このプールを IP29 と呼ぶことにする．

表 2 IP29 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	29.000	0.000	29.000	29.000
PSD	1002	0.381	0.042	0.287	0.532
能力値	1002	0.000	1.000	−2.606	2.970
推定値	1002	−0.362	0.948	−2.597	2.307

能力値と推定値の相関 0.999

このプールは 29 個のテストレットのみを含み，項目プールとしては満足できないほど小さいので，同じテストレットをそれぞれ 2 個ずつ含むプール(IP58)，3 個ずつ含むプール(IP87)，4 個ずつ含むプール(IP116)も構成した．

4.3 受験者のグループ化

石塚他(2002)では，IP58，IP87，IP116 で同じ問題が繰り返し提示されたとき，同じ解答を繰り返すデザインが採用されたが，ここでは受験者を 80 の能力群に分類し，繰り返しの提示に対しては同じ群に属する異なる受験者の解答パタンを採用することで，同じ問題に同じ解答が繰り返されることを避けようとした．

受験者の分類は，1002 名の受験者を能力の高い順に並べ，上から順に 12 名あるいは 13 名の受験者を選んで 1 群とした．具体的には，奇数番目の能力群には 13 名を割り当て，偶数番目の能力群には 12 名を割り当てた．最後の第 80 群には残りの 14 名を割り当てた．

4.4 結　果

実験 I の結果を表 2 から表 5 に示す．IP29 では予想どおり項目プールの底をつくまでに試験を終了する受験者はいなかった．これは，PARSCALE (Muraki and Bock,1996)による能力の推定においても，全て(1002 人)の受験者の PSD が 0.3162 以上であったので当然の結果である．これは，元々この試験に含まれる項目の識別力がそれほど高くないことにも起因する．

表 3　IP58 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	25.740	15.130	5.000	58.000
PSD	1002	0.320	0.018	0.289	0.422
能力値	1002	0.000	1.000	−2.606	2.970
推定値	1002	−0.370	1.014	−2.674	2.599

能力値と推定値の相関 0.986

表 4　IP87 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	18.277	16.221	7.000	87.000
PSD	1002	0.313	0.006	0.301	0.385
能力値	1002	0.000	1.000	−2.606	2.970
推定値	1002	−0.375	1.043	−2.608	2.595

能力値と推定値の相関 0.971

表 5　IP116 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	13.814	8.997	7.000	116.000
PSD	1002	0.312	0.003	0.302	0.316
能力値	1002	0.000	1.000	−2.606	2.970
推定値	1002	−0.371	1.058	−2.588	2.584

能力値と推定値の相関 0.963

このプールの結果の特徴は，能力の推定値と本来の能力値(PARSCALE で推定したもの．この値自体推定値であるが，ここでは当面興味の対象となっている CBT システムの外部で決まったものということで本来という言葉を用いる）との一致の程度が高いことである．全く同じテストレットに全く同様に解答しているので完全に一致するはずであるが，PARSCALE の能力値が 1002 人の集団で推定された項目パラメタを用いたものであるのに対し，CBT では石塚他(2001)で推定されたものをプールのパラメタとしているので微妙にずれている．その結果，本来の推定時には全員 0.3162 以上の PSD であったが，それを下回る受験者も生じている．

それでも，相関係数にして 0.999 というのは良い一致と言えよう．両者の散布図を図 2 に示す．散布図をみると，受験者の能力は本来の能力より低めに推定されるようである．これは，本来の能力値とここで言っているものが，この 1002 名の集団を対象に PARSCALE でパラメタを推定し，この集団において平均 0，標準偏差 1 に規準化したものであるのに対し，能力の推定値は CBT のパラメタを利用して推定されたものであることによると思われる．CBT のパラメタを推定した集団が 5 教科受験者の集団で学力の高い集団であったのに

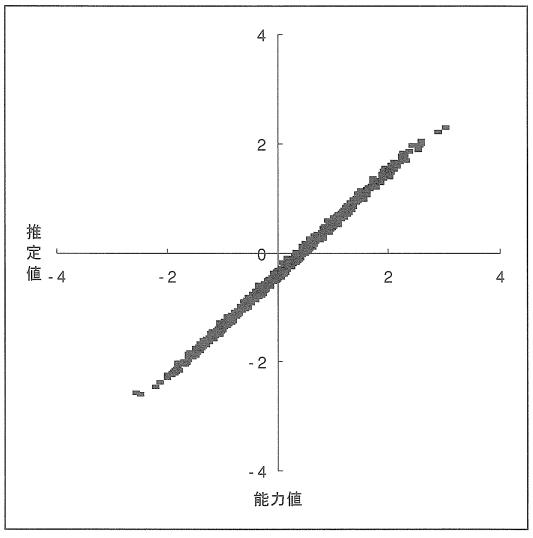

図 2　能力値と推定値の散布図(IP29)

対して，今回利用した1002名が全英語受験者からの無作為標本であるため，学力の高い集団で規準化された尺度に合わせると低めになるのだろう．

同じテストレットを2個ずつ含むIP58になると，プールの底まで行きつく受験者は相変わらず多いものの，一番早いもので7回というように，途中で試験を終了する受験者も出始める(能力値が中程度以下の者が多い)．これは，プールの大きさが2倍になった効果で，情報量表の始めの方に情報量の大きいテストレットが集まった結果である．すなわち，受験者はその能力値で情報量の大きいテストレットに2回反応をするので，見かけの精度が向上してくる．一方，元の29個のテストレットの全てに答えることにはならないので，CBTによる能力推定値は29個のテストレットで推定した本来の能力値から乖離し始める．この傾向は能力値の低い部分で大きいようだ(図3)．相関係数も0.986に下がる．

IP58で生じた効果は，IP87，IP116でさらに強まり，最小で7回，平均でも14回で試験を終了するようになる(IP116)．一方，プールの底まで到達してしまう者も相変わらず存在する．しかし，標準偏差の動向からみると，IP116では長く掛かったり，プールの底まで到達してしまう受験者の数は少ない．

今回の結果は，試験に掛かる回数が石塚他(2002)に比べて若干多めになっているが，能力推定値の平均や標準偏差はよく似た傾向を示している．大きな違いは相関の低下の仕方である．石塚他(2002)ではIP58，IP87，IP116と進むにつれて，IP29では0.999あった相関が0.971，0.922，0.879と低下したが，今回は0.986，0.971，0.963とその低下の仕方が小さい．一見同じ解答を繰り返す方が相関が高くなるように思われるが，早く試験を終える受験者は29個のテストレット全てに答えていないことを考慮する必要がある．同じ解答を繰り返して早く試験を終えるのではなく，同じ能力群に属する異なる受験者の反応を利用したことで，1回か2回解答が余計に掛かり，その結果本来の能力との一致がよくなったものと考えられる．

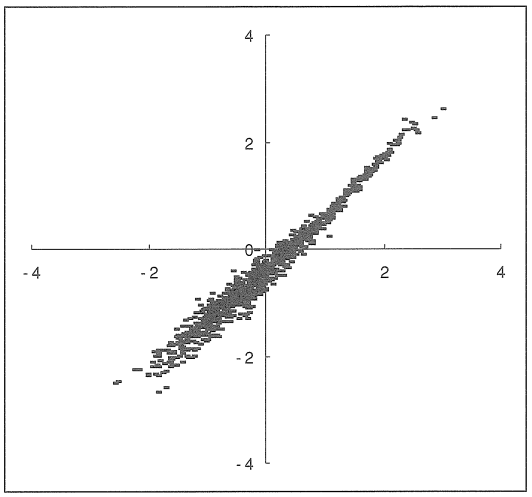

図 3　能力値と推定値の散布図(IP58)

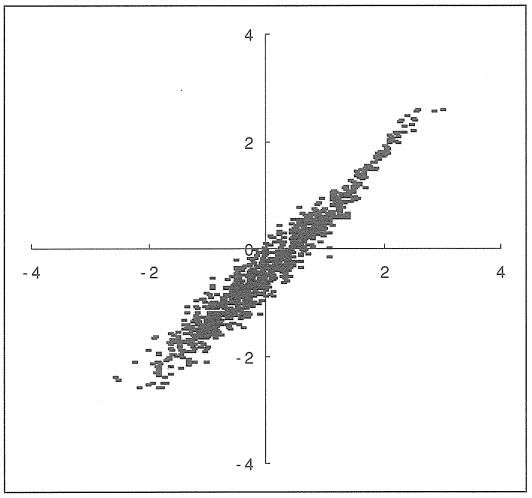

図 4　能力値と推定値の散布図(IP87)

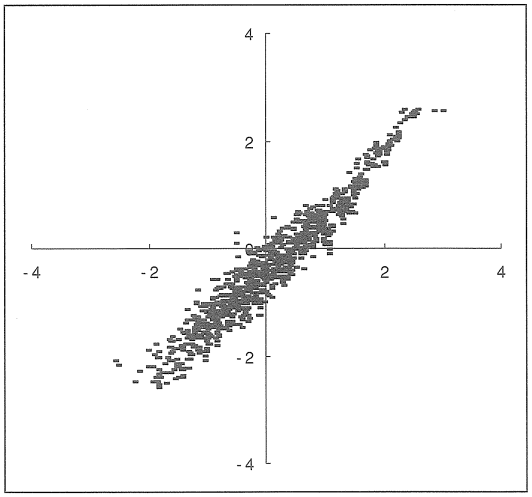

図 5　能力値と推定値の散布図(IP116)

5 確率的に生成された項目プールを用いたシステムの評価(実験II)

5.1 項目プールの生成

4章で用いた項目プールは，たとえ大きさを大きくしても同じテストレットの繰り返しになるため，その大きさには限界がある．そこでテストレットを確率的に発生させて項目プールを生成することにした．このとき識別力の高い(0.5～1.5)テストレットを含むプールと，比較的低い(0.1～1.1)テストレットを含むプールを生成することとし，それぞれを2例ずつ，また，プールの大きさも400と800で比べてみることとした．こうして

大きさ400のプール
識別力の高いもの(IP400，IP401)
識別力の低いもの(IP402，IP403)
大きさ800のプール
識別力の高いもの(IP800，IP801)
識別力の低いもの(IP802，IP803)

計8つの項目プールを生成した．

5.2 項目プール生成のアルゴリズム

1. テストレットの大きさの決定
   1～0の一様乱数を十倍して整数化し，それが2～9の範囲に収まるものを採用する．
2. 識別力の決定
   1～0の一様乱数に0.5(高識別力条件)もしくは0.1(低識別力条件)を加える．
3. 困難度の決定
   b_1 をN(0,1)の乱数から生成．テストレットの大きさが2の場合は生成された困難度を採用する．テストレットの大きさが3以上の場合は負の b_1 が得られるまで乱数を取り続ける．0.5～1の間の一様乱数 d を発生させ，$b_{k+1} = b_k + d$ とする．b_{k+1} が4を越えるときは b_1 を取り直す．

5.3 受験者の生成

実験Iの場合と異なり実受験者がいないので，受験者の能力値をN(0,1)の乱数から生成する．

5.4 解答の生成

受験者の能力値とテストレット応答関数から各カテゴリの応答確率を計算し，1～0の一様乱数を利用して解答を生成する．

5.5 結　果

実験IIの結果を表6から表13に示す．識別力の高いテストレットでプールを構成したIP400とIP401では平均5.5回程度，最大でも11回程度で試験を終了する．これは一見少ないようだが，1つのテストレットが平均5.5個の小問を含むので，平均的な受験者としては，6問程度の大問を含む試験を受験したことに相当する．

識別力の低いテストレットから構成したIP402とIP403では平均で9問程度，最大で15問程度

表6 IP400の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	5.449	0.905	4.000	11.000
PSD	1002	0.301	0.010	0.250	0.316
能力値	1002	−0.031	0.995	−2.960	2.744
推定値	1002	−0.020	0.949	−2.624	2.716

能力値と推定値の相関 0.956

表7 IP401の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	5.538	0.871	3.000	10.000
PSD	1002	0.301	0.009	0.260	0.316
能力値	1002	0.008	1.023	−4.273	3.420
推定値	1002	0.021	0.961	−2.615	2.724

能力値と推定値の相関 0.954

表8 IP402の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	9.358	1.389	4.000	15.000
PSD	1002	0.308	0.006	0.290	0.316
能力値	1002	0.008	1.015	−3.259	3.151
推定値	1002	0.013	0.960	−2.603	2.675

能力値と推定値の相関 0.954

表9 IP403の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	9.314	1.281	5.000	14.000
PSD	1002	0.307	0.006	0.295	0.316
能力値	1002	−0.038	0.967	−2.988	3.226
推定値	1002	−0.036	0.915	−2.637	2.569

能力値と推定値の相関 0.950

表 10 IP800 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	5.470	0.931	4.000	9.000
PSD	1002	0.301	0.009	0.269	0.316
能力値	1002	−0.038	0.997	−3.264	3.144
推定値	1002	−0.028	0.943	−2.614	2.621

能力値と推定値の相関 0.956

表 11 IP801 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	5.479	0.867	4.000	9.000
PSD	1002	0.300	0.010	0.278	0.316
能力値	1002	0.027	1.023	−3.491	3.151
推定値	1002	0.033	0.968	−2.599	2.659

能力値と推定値の相関 0.955

表 12 IP802 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	9.221	1.264	4.000	14.000
PSD	1002	0.307	0.006	0.292	0.316
能力値	1002	0.022	1.024	−3.143	3.122
推定値	1002	0.026	0.965	−2.587	2.629

能力値と推定値の相関 0.952

表 13 IP803 の結果

変数	観測数	平均	標準偏差	最小値	最大値
試行数	1002	9.231	1.240	5.000	14.000
PSD	1002	0.307	0.006	0.271	0.316
能力値	1002	−0.011	1.016	−3.399	2.905
推定値	1002	−0.006	0.956	−2.600	2.646

能力値と推定値の相関 0.952

を要している．一方，相関係数はあまり変わらず，0.95 程度である．古典的テスト理論にならって測定値と真値の間の相関係数の二乗をもって信頼性の指標と考えれば，0.95 の二乗は 0.9025 なので，PSD 0.3162 を試験終了の目安とすることで担保しようとした精度が保たれていることが分かる．一方，標準偏差や範囲に着目すると，推定値の散らばりは本来の能力値の散らばりよりも小さめになる傾向が見受けられる．その理由は不明だが，実受験者の解答を利用した実験 I ではなかったことである．理論的な考察を必要としよう．

プールの大きさを 2 倍にした IP800 から IP803 の結果もほぼ同様で，400 もあれば，測定の見地から言ってプールを大きくする効果はあまりないようだ．勿論，セキュリティの見地から言えば，プールは大きいに越したことはない．ただ，システム

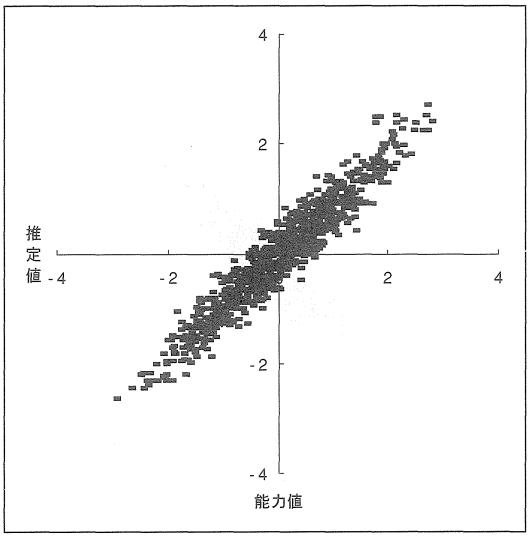

図 6 能力値と推定値の散布図(IP400)

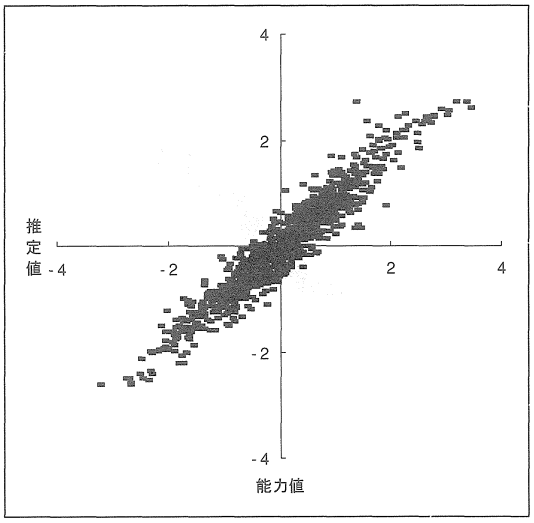

図 7 能力値と推定値の散布図(IP401)

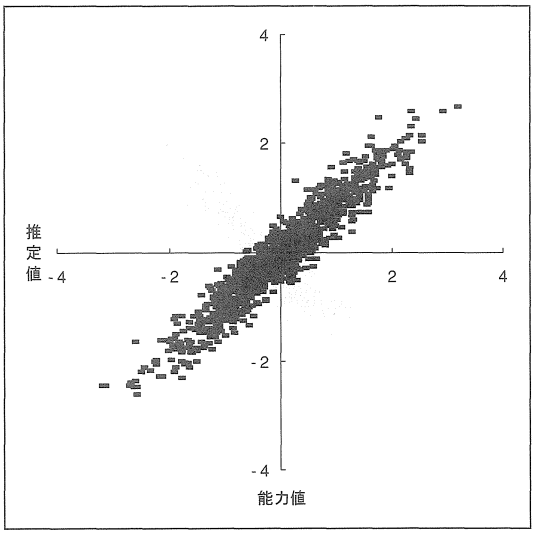

図 8 能力値と推定値の散布図(IP402)

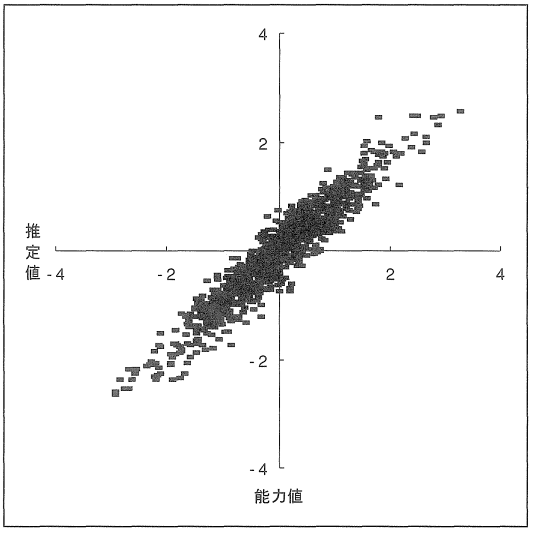

図 9 能力値と推定値の散布図(IP403)

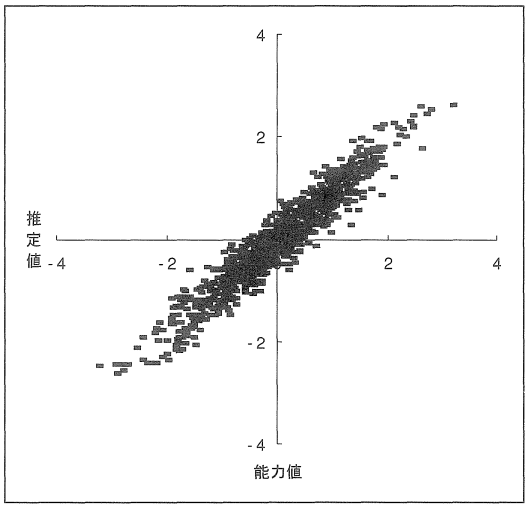

図 10 能力値と推定値の散布図(IP800)

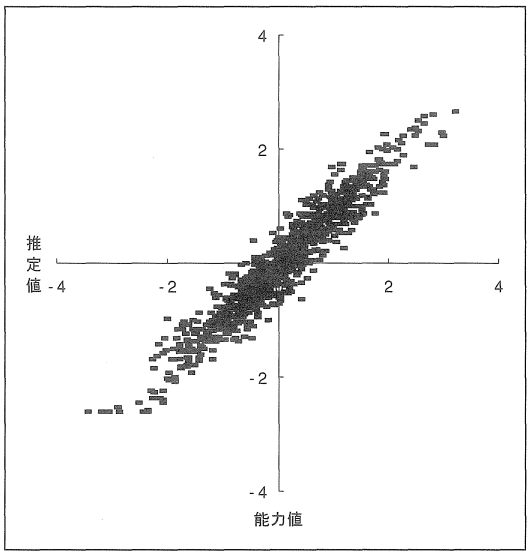

図 11 能力値と推定値の散布図(IP801)

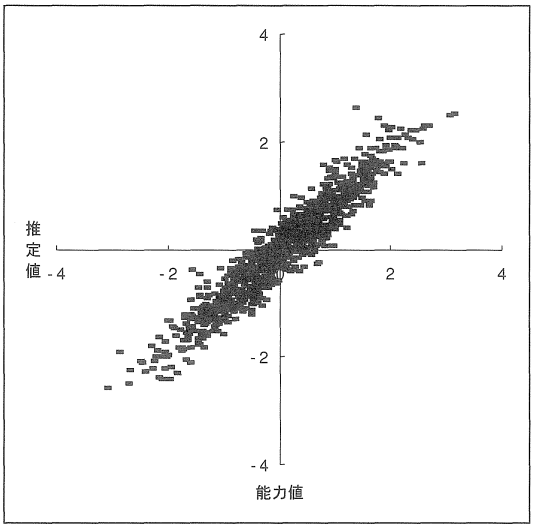

図 12 能力値と推定値の散布図(IP802)

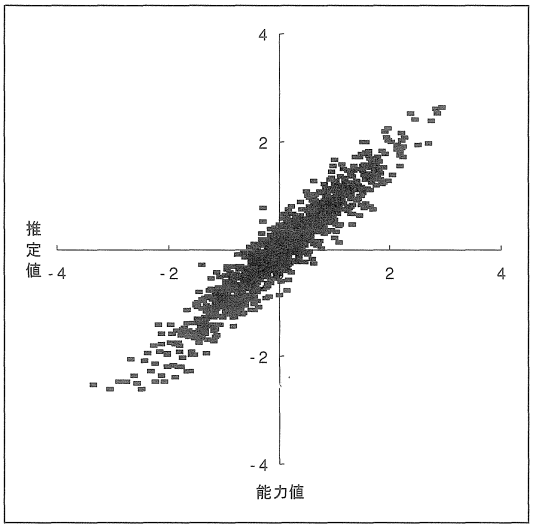

図 13 能力値と推定値の散布図(IP803)

の立ち上げ時に，最低限どの程度の大きさのプールを用意する必要があるのかを明らかにするために，今後はより小さいプールを構成して実験して行くことも意味があるだろう．

6 おわりに

実受験者が存在する項目プール(実験 I)と仮想的な項目プール(実験 II)によって CBT システムの評価を試みた．実験 II によれば CBT システムはほぼ期待通りに動くように思われる．ただ，プールも解答も確率的に発生させる実験 II はご都合主義の感を免れない．是非とも実受験者の存在する

項目プールを用いてシステムの評価を行いたいところだが，センター試験の範囲で考える限り共通受験者の存在する問題数はあまり多くない．実験Iでは同じ能力群に属する異なる受験者の解答の採用によって石塚他(2002)よりも妥当性を高めているが，実受験者の存在する大きな項目プールを採用するのが本来の有り方だろう．1つの方法として，1科目だけではなく，英語と共通受験者の存在する国語や公民，地歴などの科目も含めて文系の試験と考え，項目プールを構成することが考えられるが，1次元性に疑問が残り，あえてそれを無視しても得られる問題の数がそれほど多くないので，逡巡しているところである．

参考文献

石塚智一・中畝菜穂子・内田照久・前川眞一 (2001) テストレットモデルによる英語試験問題の分析，大学入試センター研究紀要，30，21–38.

石塚智一・前川眞一・菊地賢一・中畝菜穂子・内田照久 (2002) テストレットモデルによるCBTシステムの構築とその評価，大学入試センター研究開発部リサーチノート，RN-01-11.

Linn, Robert l., Rock, Donald A., and Cleary, T. Anne (1969) The development and evaluation of several programmed testing methods, *Educational and Psychological Measurement*, 29, 129–146.

Lord, Frederick M.(1970) Some test theory for tailored testing. In W. H. Holtzman (Ed.) *Computer-assisted instruction, testing, and guidance* (pp139–183) New York: Harper & Row.

Muraki, Eiji and Bock, R. Darrel (1996) *PARSCALE™ IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*, Version 3, IL: Scientific Software International.

Rosenbaum, Paul R. (1988) Item bundles, *Psychometrika*, 53, 3. 349–359.

Samejima, Fumiko (1969) Estimation of latent trait ability using a response pattern of graded scores, *Psychometric Monograph*, 17, 1969.

Samejima, Fumiko (1972) A general model for free response data, *Psychometric Monograph*, 18, 1972.

Thissen, David, Steinberg, Lynne, and Mooney, Jo Ann (1989) Trace lines for testlets: A use of multiple—categorical—response model, *Journal of Educational Measurement*, 26, 3, 241–260.

Thissen, David and Mislevy, Robert J. (2000) Testing algorithms. In Howard Wainer (Ed.) *Computerized adaptive testing: A primer*, Second edition (pp101–133) New Jersey: Lawrence Erlbaum Associates, Publishers.

Wainer, Howard, and Kiely, Gerald L. (1987) Item clusters and computerized adaptive testing: A case of Testlets, *Journal of Educational Measurement*, 24, 3, 185–201.

Computerized Adaptive Testing System Based on Testlet Response Model

ISHIZUKA Tomoichi*
MAYEKAWA Shin-ichi**
KIKUCHI Ken-ichi***
NAKAUNE Naoko*
UCHIDA Teruhisa****

Abstract

A computerized adaptive testing system which is based on testlet response model was developed.

The system was evaluated through two procedures: one using real examinee' s response patterns, and the other using generated ability scores.

Very high correlations between the true abilities and the estimated abilities were found through the first evaluation, and, through the latter procedure, it was confirmed that the system works fairy well.

Key words: computerized adaptive testing, item response theory, testlet response model.

* Department of Applied Statistics and Measurement, Research Division, The National Center for University Entrance Examinations, 2-19-23 Komaba, Meguro-ku, Tokyo 153-8501 Japan
** Graduate School of Decision Science and Technology, Tokyo Institute of Technology
*** Faculty of Science, The Toho University
**** Department of Comprehensive Studies on Admission and its Environment, Research Division, The National Center for University Entrance Examinations