Bulletin of the Graduate School of Education and Human Development, Nagoya University (Psychology and Human Development Sciences) 2002, Vol. 49, 47 − 54.

語彙理解尺度におけるCBT版と紙筆版の同等性の検証

―― 項目反応理論によるテスト作成・分析を通した検討 ――

熊　谷　龍　一[1)]

はじめに

昨今のコンピュータ資源の普及により，心理測定，能力測定に際してコンピュータを用いたテスト（Computer Based Test: CBT）が行われるようになってきた（廣瀬，2000）。テスト媒体が紙と鉛筆からコンピュータに変わることは，ただ単に測定道具が変化したということにとどまらない。紙筆によるテストでは行うことができなかったテスト方式をCBTにより行うことができる。コンピュータを利用することにより実用化されることになったテスト方式の一つとして，項目可変型の適応型テスト（adaptive test）が挙げられる。

適応型テストは，テストの各受験者の能力を適宜推定しながら，その受験者に最も適した項目を提示していくものである。この方法によって各受験者に提示される項目は受験者ごとに異なるものとなるが，項目反応理論による尺度化によってそれらを共通の尺度上で比較することが可能となる。また項目固定型のテストに比べて，少数の項目で精度の良いテストを実施することができる。実際の適用例としては，外国語としての英語能力試験の一つであるTOEFLにおいてコンピュータによる適応型テストが採用され，実際に運用されている。国内でも，柴山・野口・芝・鎌原（1987）や，服部（1990）などにより，日本語の語彙理解力の測定を目的とした適応型テストの開発，検証が行われている。

コンピュータによるテストを作成する場合には，項目作成の時点からすべてを新規に作成する場合と，これまでの既存の尺度をCBT化する場合とがある。

先に挙げた柴山他（1987）は，芝（1978）による語彙理解尺度を適応型CBTとしたものである。ここで問題となるのは，CBTとこれまでの紙筆版テストの同等性である。尺度をCBT化することにより，紙筆版テストとは異なる部分が生じる。それは入出力装置の違いであったり，テストの形式の違いであったりする。テスト形式の違いとは，例えば適応型テストにおいて一度解答した項目については，他の項目に解答すると再び解答し直すことができないなどである。当然紙筆テストでは時間内であれば，冊子内を移動することで以前の問題を解答し直すことができる。

既存の尺度をCBT化する際には，これらの相違を踏まえた上で，CBTが紙筆テストと同等（Equivalent）であることを確認することが重要となる（APA, 1986）。

本研究では，芝（1978）による語彙理解尺度をCBT化したものについて，紙筆版との同等性が保たれていることを検討する。語彙理解尺度をCBT化したものとしては，先に挙げた柴山他（1987）によるものがあるが，これは適応型テスト方式によるものである。適応型テストでは，受験者に提示される項目および項目数が個人ごとに異なる。紙筆テストでは，全受験者に対して共通の項目が用いられるため，この両者を直接比較するのが困難となる。そこで本研究では，項目を固定型としたCBT版語彙理解尺度を作成し，紙筆版と比較検討することによって同等性を確認する。また，本研究ではテスト構成・分析に項目反応理論を用いた。本研究では同等性の確認のために，芝（1978）による語彙理解尺度の項目群の中から受験者の能力水準に合った困難度を持つテストを2つ作成するという手続きをとった。

CBT版語彙理解尺度の開発

今回語彙理解尺度をCBT化するにあたっては，Microsoft VisualBasic 6.0を用いてPC版語彙理解尺度を開発した。

実際のテスト画面

実際のテスト画面はFigure 1のようなものである。画面左側に残り問題数が表示され，画面右側に問題項目と解答選択肢が提示されている。被験者は問題欄に提示されている語句と最も意味の近い言葉を，画面右の解答選択肢欄から選び，マウスで選択する。選択された選択肢は，目立つように黄色で表示される。選択が終わると，被験者は次の問題へ進むために「次の問題へ」と書かれ

1）名古屋大学大学院教育発達科学研究科博士課程（後期課程）

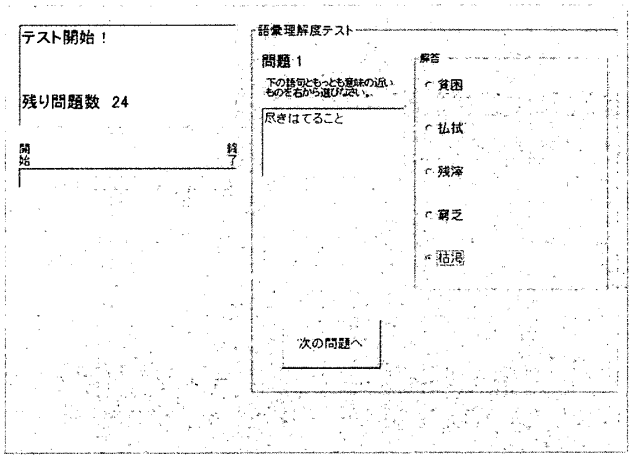

Figure 1　テスト画面

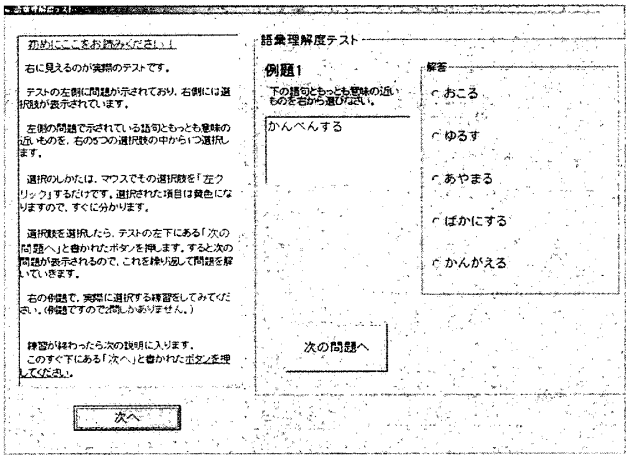

Figure 3　チュートリアル画面

たボタンを押す。このボタンが押されると，次の問題項目が提示される。この手続きを繰り返す形で，テストは進行していく。

また，今回作成したCBTが紙筆テストと大きく異なる側面として，一度解答した項目に戻れないことが挙げられる。そこで，もし被験者がなにも選択せずに次の項目へ進もうとした時には，警告音とともにFigure 2のようなダイアログが表示され，本当に未選択のまま次の問題へ進んでよいかを確認するようにした。

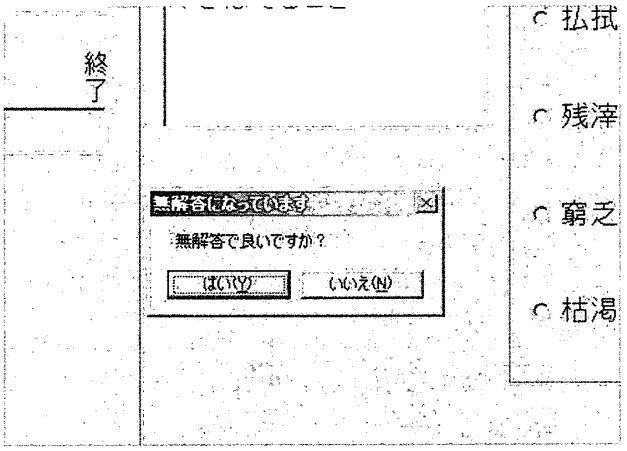

Figure 2　警告ダイアログ

チュートリアル

CBTは画面に提示される問題項目の形式や解答するための操作方法など紙筆テストと異なる点が多いため，被験者にとって初めての経験となる事柄が多い。そこで被験者が円滑にCBTでの解答ができることを目的としたチュートリアルを作成し，テストの前に被験者に提示するようにした。

画面の見方の確認　今回作成したCBTは紙筆版テストとは画面の構成が異なる。具体的には，紙筆版は冊子一枚あたりに複数の項目が提示されているのに対して，CBT版は一画面に一つの項目が提示され，それに解答した後に次の項目が提示される形となる。したがって，CBT版の被験者にはテストの画面構成がどのようになっているのかを示すようなチュートリアルを作成し，テスト開始前に被験者に提示した（Figure 3参照）。これにより，被験者はどこに問題や選択肢が提示され，テストはどのように進んでいくのかを確認することができる。

操作方法の確認　紙筆版テストとのもう一つの違いが，入力方法である。CBT版では鉛筆の代わりにマウスが入力手段となるため，被験者がマウス操作によりテストに解答するということを認識する必要がある。そのため，チュートリアル内でマウスを使ってテストに解答するという説明がなされるとともに，2つの例題に実際にマウスを用いて解答する経験をもつという過程を導入した。

尺度の同等性の検証

問題

本研究ではCBT版と紙筆版の語彙理解尺度の同等性を検証するのに，大学生を被験者として用いる。APA (1986) では，CBTと紙筆版テストの同等性を検証するために，両形式に解答した被験者の得点順位がほぼ対応していること，平均・分散・得点分布の形がほぼ等しいか，尺度変換すれば等しくなること，が求められている。しかし，実際には両形式のテストには同一の項目が含まれていることがあるので，同一被験者が両方のテストを受験すると同じ項目を2度解答することになる。このことを避けるために，今回の研究ではCBT版と紙筆版による比較テストと，紙筆版のみですべての被験者が受験する共通テストという2つのテストを作成した。（Figure 4参照）

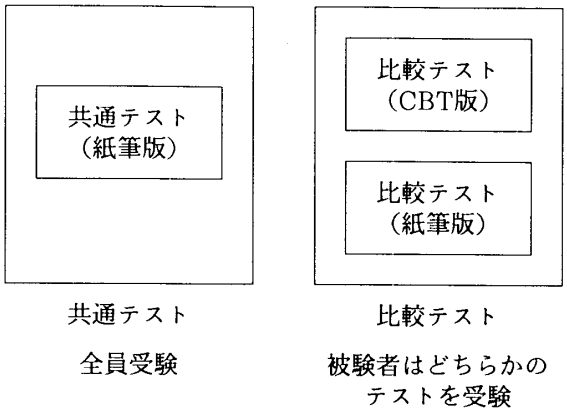

Figure 4　テストデザイン

テストの構成

芝（1978）の語彙理解尺度は，小学1年生から大学生までを対象とした11版の下位尺度で構成されている。後に芝・野口（1982）によって，これらが等化され，全ての項目のパラメタ値が共通尺度上で表された単一の項目プールの形で整備されている。芝（1978），芝・野口（1982）では項目反応モデルとして2パラメタ・ロジスティック・モデルが採用されている。2パラメタ・ロジスティック・モデルでは項目jの特性曲線は，

$$P_j(\theta)=\frac{1}{1+\exp[-1.7a_j(\theta-b_j)]} \qquad (1)$$

で表され，変数θの関数である。θは被験者の能力（ここでは語彙理解力）を示すものである。このモデルでは項目の特徴を示すパラメタとして，識別力を表すパラメタaと，困難度を表すパラメタbが導入されている。芝・野口（1982）の項目プールには，この2つのパラメタ情報が含まれている。項目反応理論では，このような整備された項目プールの情報を利用することで，予備テストを実施することなしに，被験者に最適なテストを構成することが可能となる。具体的には，先の識別力・困難度パラメタ，さらにそこから算出されるテスト情報量をもとにテストを構成することになる。テスト情報量は，真の能力値θに対する最尤推定量$\hat{\theta}$の分散の逆数であり，テストの精度を示している。2パラメタ・ロジスティックモデルにおけるテスト情報量は，

$$I(\theta)=1.7^2\sum_{j=1}^{n}a_j^2P_j(\theta)\{1-P_j(\theta)\} \qquad (2)$$

で算出される。ここでnはテストに含まれる項目数である。テスト情報量も能力値θの関数になっていることから分かるように，項目反応理論ではテストの精度に関する情報を被験者の能力値ごとに算出することができる。

本研究では大学生を被験者とすることから，以下のような基準を満たす共通テストおよび比較テストの項目を各テストにそれぞれ25項目ずつ選択した。

1．両テストともに困難度パラメタの平均値が4.0前後であること[2)]。
2．両テストの識別力，および困難度パラメタの平均値・標準偏差がほぼ等しくなること。
3．両テストの項目パラメタから計算されるテスト情報関数の形状がほぼ等しくなること。

共通テストおよび比較テストに含まれる項目の識別力・困難度パラメタをTable 1に示す。また両テストのテスト情報関数をFigure 5に示す。

Table 1　共通テストおよび比較テストの項目パラメタ

共通テスト			比較テスト		
項目	識別力	困難度	項目	識別力	困難度
1	0.4	2.5	1	0.4	2.5
2	0.5	2.8	2	0.5	2.9
3	0.4	3.0	3	0.8	3.0
4	0.4	3.0	4	0.3	3.0
5	0.3	3.0	5	0.3	3.0
6	0.6	3.1	6	0.6	3.1
7	0.8	3.4	7	0.6	3.1
8	0.7	3.4	8	0.6	3.4
9	0.5	3.4	9	0.5	3.4
10	0.4	3.4	10	0.4	3.5
11	0.4	3.6	11	0.4	3.6
12	0.5	3.9	12	0.3	4.0
13	0.3	4.0	13	0.7	4.1
14	0.5	4.1	14	0.4	4.1
15	0.5	4.2	15	0.3	4.1
16	0.3	4.2	16	0.6	4.4
17	0.4	4.4	17	0.5	4.4
18	0.6	4.5	18	0.3	4.4
19	0.4	4.5	19	0.4	4.5
20	0.4	4.6	20	0.3	4.5
21	0.3	4.7	21	0.7	5.0
22	0.5	5.3	22	0.5	5.0
23	0.4	5.4	23	0.5	5.4
24	0.6	5.6	24	0.6	6.0
25	0.7	6.1	25	0.4	6.0
平均	0.47	4.00	平均	0.48	4.02
標準偏差	0.13	0.92	標準偏差	0.14	0.94

2）芝・野口（1982）では尺度の原点と単位について，中学1年生の特性値の平均が0.0，標準偏差1.0となるように等化なされていた。そして，高校3年生の特性値の修正平均が4.08となっているため，この値を設定した。

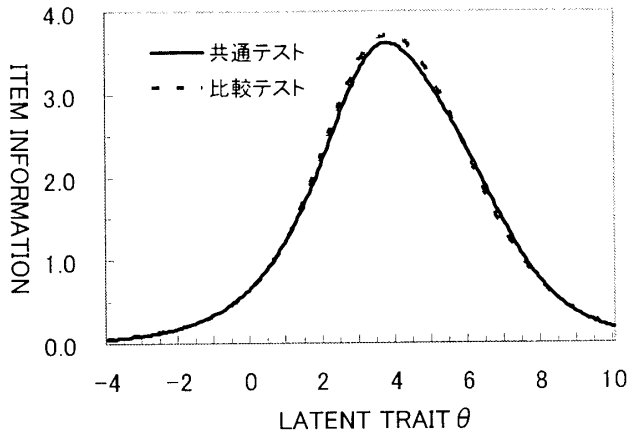

Figure 5　共通テストおよび比較テストの
テスト情報曲線

方法

被験者　被験者は名古屋市内の大学生および大学院生158名であった。そのうち共通テストとCBT版比較テストを受験したのが30名，共通テストと紙筆版比較テストを受験したのが128名であった。

装置　CBT版比較テストの実施は，NEC製ノートパソコンPC-LG93JJHEBにマウスを接続したものが用いられた。

手続き　CBT版比較テストを受験する被験者群に対しては，個別にテストを実施し，紙筆版比較テストを受験する被験者群に対しては，教室で一斉にテストを実施した。

CBT版テストを実施する前に，各被験者にマウス操作の経験を尋ねたところ，30名全ての被験者がマウス操作の経験があった。CBT実施に際しては，チュートリアルで示される指示にしたがって操作を進めていくよう指示をした。また，チュートリアル終了後に質問がある被験者は実験者に尋ねることができるようにした。

どちらの被験者群についても，テストを受ける順序による効果を消去するために，共通テストと比較テストを受験する順序は被験者ごとにランダムに実施した。

結果と考察

テストの同等性を検証するための分析は，古典的テスト理論を用いた分析と項目反応理論を用いた分析とに分けて行われた。

古典的テスト理論を用いた分析

ここでは，CBT版と紙筆版の比較テストについての分析を行う。

信頼性係数　両テストの信頼性係数（KR20）を算出したところ，Table 2のようになった。

Table 2　両テスト形式の信頼性係数

テスト形式	信頼性係数
CBT　版	.71
紙　筆　版	.63

これらの値は，能力測定テストとしてはやや低い値である。この原因としては，語彙理解力尺度の項目プールから大学生を対象としたテストを2組作成した結果，各テストが25項目ずつというやや少な目のものとなったことが挙げられる。芝（1978）では，大学生を対象とした下位尺度であるU版では，項目数が54であった。そこで共通テスト25項目を足し合わせた50項目のテストとして信頼性係数を算出すると，CBT版（＋紙筆版の共通テスト）では.86，紙筆版（＋紙筆版の共通テスト）では.80となった。もちろん，CBT版に関しては比較テストと共通テストでテスト方式が異なるので，一概に一つのテストの信頼性としてみることはできないが，紙筆版においては比較テストも共通テストも紙筆方式であるため

Table 3　CBT版および紙筆版比較テスト
の記述統計量

	CBT版テスト (N＝30)		紙筆版テスト (N＝128)	
項目番号	通過率	標準偏差	通過率	標準偏差
1	.89	0.31	.83	0.38
2	.64	0.48	.50	0.51
3	.91	0.29	.97	0.18
4	.81	0.39	.83	0.38
5	.57	0.50	.50	0.51
6	.49	0.50	.43	0.50
7	.96	0.20	1.00	0.00
8	.38	0.49	.57	0.50
9	.53	0.50	.37	0.49
10	.58	0.50	.63	0.49
11	.41	0.49	.50	0.51
12	.70	0.46	.73	0.45
13	.64	0.48	.53	0.51
14	.72	0.45	.67	0.48
15	.36	0.48	.37	0.49
16	.90	0.30	.90	0.31
17	.63	0.48	.43	0.50
18	.57	0.50	.37	0.49
19	.51	0.50	.60	0.50
20	.62	0.49	.40	0.50
21	.54	0.50	.60	0.50
22	.30	0.46	.27	0.45
23	.22	0.42	.20	0.41
24	.25	0.44	.27	0.45
25	.30	0.46	.33	0.48
	平　均	標準偏差	平　均	標準偏差
合計得点	13.80	4.00	14.43	3.55

50項目からなる一つのテストの信頼性係数としてみることができ，項目数増加により信頼性が向上することが確認できる。

テストおよび各項目の平均・標準偏差　各テスト形式における項目の通過率と標準偏差，および正答数得点の平均と標準偏差はTable 3に示してある。ここで合計得点によるt検定を行ったところ，両テスト形式の間に有意差は見られなかった（t (156) $=-0.86$, $p>.05$)。

以上の結果から，信頼性係数はやや低いものの，正答数得点からみた両テスト形式の差異は確認されなかった。

項目反応理論を用いた分析

1次元性の確認　2パラメタ・ロジステック・モデルにおいては，測定しようとしている特性が1次元であることが仮定される。そして尺度の1次元性を確認する手がかりの一つに，四分相関係数行列を主因子分析することで得られる固有値を検討する方法がある。今回利用した語彙理解尺度は芝（1978）により1次元性の確認が既になされている。本研究では，語彙理解尺度から項目を選択してCBT版と紙筆版のテストを作成したことから，

Table 4　項目間四分相関行列の固有値
（値の大きいものから10個まで）

因子数	CBT版	紙筆版
1	5.000	6.284
2	2.493	3.375
3	2.107	2.835
4	2.030	2.498
5	1.806	2.090
6	1.624	1.867
7	1.396	1.727
8	1.267	1.365
9	1.227	1.155
10	1.162	0.969

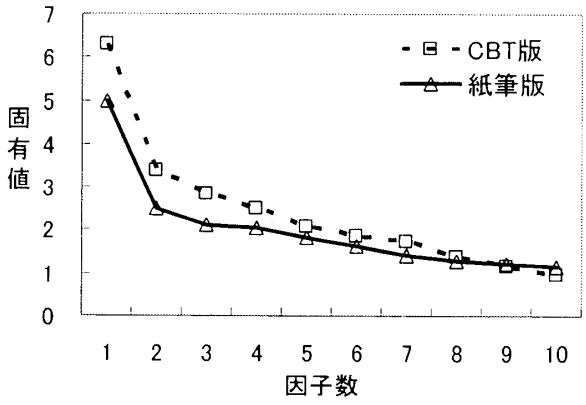

Figure 6　スクリープロット
（値の大きいものから10個まで）

それぞれが意図したとおり1次元性を示しているのかを再検討する。またCBT版のテストにおいて，測定しようとしている能力以外の要因がテスト結果に関係しているのかについて確認することができる。なお，四分相関係数の推定および因子分析にはTESTFACT（Wilson, Wood & Gibbons, 1985）を用いた。CBT版比較テストに関しては項目7が通過率1.00となったため，因子分析からは除外した。

Table 4は両テスト形式の固有値である。Figure 6はそれをグラフ化したスクリープロットである。第2から第4固有値の値もやや大きいが，スクリープロットを見ても第1固有値から第2固有値の落ち込みが他に比べて大きくなっていることから，特に2因子以上をとる根拠は無いといえる。

等化係数によるテストの同等性の検討　Figure 4に示したとおり，本研究ではすべての被験者が共通に受験する紙筆版共通テストを設定した。被験者は共通テストと比較テスト（CBT版か紙筆版のどちらか）の2つのテストを受験した。したがって各被験者に対して共通テストの潜在特性値と比較テストの潜在特性値とが算出される。

このように2つのテストから得られた特性値を比較するためにはそれらが同一の原点と単位を持つ尺度上に乗っていなければならない。項目反応理論では2つのテストの特性値を同一尺度上に乗せるために，等化係数により項目および特性値パラメタを変換することができる。本研究では，各テスト間で共通の受験者が存在するというテストデザインになっていた。そこで共通テストと比較テストを等化する方法として，野口（1983）による共通被験者を用いた等化を採用することができる。

本研究で用いた語彙理解尺度の各項目は芝・野口（1982）により既に等化がなされているため，共通テストと比較テストが既に共通の尺度に乗っている。したがって，両テストを等化した時の等化係数は$k=1.0$; $l=0.0$となる（実際には測定誤差により，多少値が変動する)。ここで等化係数が先の値から大きくずれていた場合，共通テストから得られる潜在特性値と比較テストから得られる特性値が同一尺度に乗っていなかったことになる。本研究で作成されたCBTが紙筆テストとの同等性が保たれていなかったために，CBTによる特性値の尺度と紙筆テストによる特性値の尺度がずれていた場合，紙筆版共通テストとCBT版比較テストを受けた受験者群において両テストを等化した時の等化係数が先の値からずれてくることになる。

そこでCBT版比較テストを受けた被験者群と紙筆版比較テストを受けた被験者群のそれぞれにおいて共通テ

Table 5　各テスト形式の等化係数の推定値

等化係数	CBT版	紙筆版
k	0.96	0.91
l	−0.14	0.27

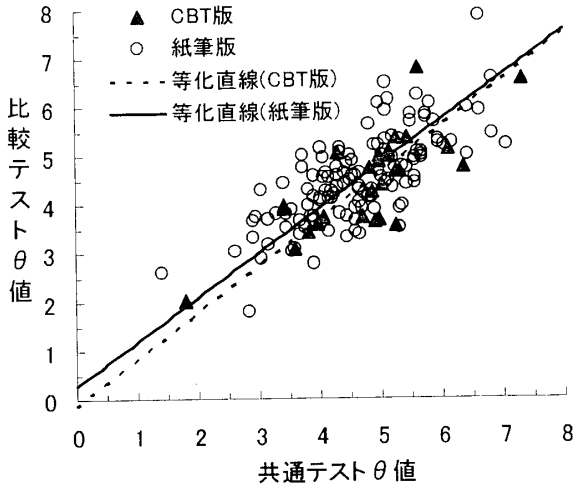

Figure 7　比較テストを共通テストに等化したときの等化直線

ストと比較テストの等化を行い，その等化係数がいずれも先の値に近ければ，版による同等性を確認する一つの根拠となる。

Table 5 は，CBT版および紙筆版の比較テストを共通テストに等化した時の等化係数である。ここで先の1次元性の確認で分析から削除したCBT版比較テストの項目7についても含めて分析が行われた。Figure 7 は等化係数をもとに，グラフを描いたものである。これらの値は，先に予測したk=1.0，l=0.0に近く，グラフからはこの2本の直線はほぼ等しいものといえる。

Table 6 は，等化後のθの値を等化前のθ値が0.00から8.00まで1刻みで記したものである。ここでFigure 5に示した両テストのテスト情報量を見ると，どちらのテストもテスト情報量の最大値は3.80より小さい。テスト情報量の逆数の平方根が，推定値$\hat{\theta}$の標準誤差となる。テスト情報量が3.80である場合の標準誤差は0.53となる。つまり両テストともに，$\hat{\theta}$の標準誤差は少なくとも0.53よりは大きくなるといえる。ここで，Table 6 の，等化前との差を見ると，どの値も絶対値で0.53よりは小さくなっている。つまり，等化前と等化後の差は標準誤差の範囲に入っている。等化前と等化後の差が誤差の範囲であるということは，等化係数がk=1.0，l=0.0から大きくずれていないと考えられる。つまりこの結果から，両テスト間で同等性があることが示唆される。

まとめ

本研究では，テストの信頼性，正答数得点の差，1次元性の分析，等化係数の分析を通して，CBT版テストと紙筆版テストの同等性を検証してきた。結果として両テスト方式の間では違いが見られず，CBT版テストと紙筆テストの同等性を確認することができた。

今後の課題

本研究では被験者が大学生および大学院生であった。現在ではほとんど全ての大学生がコンピュータを扱った経験がある。本研究でも，CBT版テストを受験した被験者は，全てがコンピュータおよびマウスによる操作を以前に経験していた。したがって本研究の分析では，コンピュータ使用の経験がCBT版のテストに与える影響を論じることはできない。しかし，小学生・中学生においてはいまだコンピュータを使用した経験がない児童・生徒も多く存在する。彼らに対して，CBT版のテストが紙筆版のテストと同等性を持ちうるのかを調査することは，今後の課題となる。

また，本研究ではCBT版を受験した被験者数は30名であった。しかしCBT版，紙筆版ともにより多くの被験者を対象とし，項目パラメタを再推定しそれらを比較したり，DIF（Differential Item Functioning）分析などからより詳細な同等性の検証を行うことが必要である。

引　用　文　献

American Psychological Association 1986 *Guidelines for computer-based tests and interpretations*. Washington DC: Author.

服部　環　1990　個人差に応じたテスト方式による語彙

Table 6　等化前と等化後のθの差

	等化前θ	0.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00
CBT版	等化後θ	−0.14	0.82	1.78	2.74	3.70	4.66	5.62	6.58	7.54
	等化前との差	0.14	0.18	0.22	0.26	0.30	0.34	0.38	0.42	0.46
紙筆版	等化後θ	0.27	1.18	2.09	3.00	3.91	4.82	5.73	6.64	7.55
	等化前との差	−0.27	−0.18	−0.09	0.00	0.09	0.18	0.27	0.36	0.45

理解力の測定　教育心理学研究，**38**，445-454.
廣瀬英子　2000　心理測定尺度のコンピュータ・テスト化に向けての最近の動向　教育心理学研究，**48**，235-246.
野口裕之　1983　被験者の推定尺度値を利用した潜在特性尺度等化法　教育心理学研究，**31**，233-238.
芝　祐順　1978　語彙理解尺度作成の試み　東京大学教育学部紀要，**17**，47-58.
芝　祐順・野口裕之　1982　語彙理解力尺度の研究　I－追跡データによる等化－　東京大学教育学部紀要，**22**，31-42.
柴山　直・野口裕之・芝　祐順・鎌原雅彦　1987　最適化テスト方式による語彙理解力の測定　教育心理学研究，**35**，363-367.
Wilson, D. T., Wood, R., & Gibbons, R. 1985 *TESTFACT*. Chicago, IL: Scientific Software International.

謝辞

本研究を実施するに当たって，語彙理解力尺度の利用を許可してくださいました東京大学名誉教授　芝　祐順教授に心から感謝いたします。また論文の執筆に当たりご指導いただきました名古屋大学大学院教育発達科学研究科　野口裕之教授，ならびに実験に多大な協力をいただいた名古屋大学大学院教育発達科学研究科　脇田貴文氏に深く感謝いたします。

（2002年9月30日　受稿）

ABSTRACT

An investigation of equivalence between CBT and Paper Pencil test

Ryuichi KUMAGAI

The purpose of this research was investigating the equivalence between computer based test version and paper pencil test version in scale for acquisition of Japanese word meanings. CBT version was developed for this study newly. Item response theory was used for constructing scaling test items and analyzing subjects item response data. These tests were administered to 158 college students and graduate students in Nagoya city (CBT = 30, paper pencil test = 128).

It was concluded that CBT version and a paper pencil test version are equivalent. It is required more study to investigate the influence of the experience operating the computer.

Key words: computer based test, paper pencil test, equivalence, item response theory, scale for acquisition of Japanese word meanings