200400960A

厚生労働科学研究研究費補助金
医療技術評価総合研究事業

医師国家試験コンピュータ化に関する研究

平成１６年度　総括・分担研究報告書

主任研究者　細田 瑳一

平成 １７（２００５）年 ４ 月

目　　次

厚生労働科学研究費補助金（医療技術評価総合研究事業）

総括研究報告書

医師国家試験コンピュータ化に関する研究

主任研究者　細田 瑳一　財団法人日本心臓血圧研究振興会附属榊原記念病院病院長
高林克日己　千葉大学医学部附属病院企画情報部教授

研究要旨

【目的】国家試験にコンピュータを利用した試験を導入する蓋然性を検討する。【方法】MCQ 問題 100 題と患者管理問題（PMP）５題、及び PMP 問題に対応する内容の MCQ 問題 10 題を作成し、６つの大学の医学生５、６年生を２群に分けて従来のマークシート法（MAR）とコンピュータ試験（COM）で半分ずつを行い、全ての結果を総括し解析することで COM と MAR の相関と利点欠点を検討する。【成績】COM と MAR の MCQ に関する解答の結果は近似しており、試験運用上でも大きな問題はなかった。PMP の COM の結果と MCQ50 題の正答率との間には評価項目及び解析可能の内容が異なるので強い相関はみられなかった。また PMP に対応する MCQ 問題とも相関が弱く、PMP での評価判定は知識のファクターだけではないことが推測された。また PMP 問題ではとくに能力の低い学生の識別が明瞭であった。しかし多数の受験生の同時受験などでの運用上の問題が解決されていないことと、学生自身は COM に賛成する者より反対意見を述べる者が MCQ で 82:381、 PMP で 116:402 と圧倒的に多かった。決定的に有用な事由がないと早急にコンピューター試験を導入することには困難があると考えられるが、試験自身には問題のないことが示されたことから、CAT の導入が可能になるなどさまざまな利点を考慮して、欠点を克服した段階で近い将来導入されることは十分考えられる。【結論】試験方法としては国家試験にコンピュータを導入し、従来のマークシートに変わり共用試験の CBT のように試験ができること、PMP のようにペーパーテストで検定できない能力の判定ができる可能性を示した。

分担研究者　高林克日己
千葉大学医学部付属病院
医療情報部教授

吉岡俊正
東京女子医科大学
医学教育学　教授

福井次矢
聖路加国際病院院長

江口光興
獨協医科大学
小児科学教授

杉山幸比古
自治医科大学
呼吸器内科学教授

山内俊雄
埼玉医科大学
精神医学教授

福島　統
東京慈恵会医科大学
解剖学　教授

原田研介
日本大学医学部
小児科学教授

分担協力者　椎橋実智男
埼玉医科大学
医療情報施設　助教授

宮木浩行
三菱電機インフォメーションシステムズ株式会社

Ａ．研究目的

コンピュータのMCQ方式と従来のマークシート試験の間にどのような差があるかの検定と、PMP方式の問題と従来のマークシートによるMCQ問題との間にどの程度の相関があるかを検定し、運用についての問題点と受験生の対応、利点欠点を示すと共に検討し、コンピュータ試験実施の蓋然性を明らかにすることを目的とする。

Ｂ．研究方法

１　問題作成

昨年度までに作成したCBTに準じたコンピュータでの multiple choice question(MCQ)のツールを用いて、各大学が独自に作成した５、６年生の内科系、外科系問題の中から同一の分野でレベルの近似した問題を集めてそれぞれ対の問題になるような100題を抽出し、各Ａ、Ｂ問題50題とし、それぞれの問題をコンピュータ用とペーパー試験用に使えるよう作成した。PMPは昨年までに作成したPMPのツールを利用してコンピュータ用に５題を高林が作成し、このうち４題については各PMP問題に対応するようにPMP１題あたり５問のMCQ問題をペーパー試験用に作成した。

２　試験方法

今回は６大学において主に５学年の学生を対象に試験を行った。試験はコンピュータによるMCQ試験50題(75分)、PMP試験３題(45分)と、ペーパー試験のMCQ60題(75分)である。各施設ではグループを対の２群（グループI、グループII）に分け、グループＩはコンピュータでMCQのＡ問題の試験、PMPでは３問を受験し、ペーパー試験ではMCQのＢ問題、およびPMPで受けなかった２問に対応するペーパー試験10題を受験した。グループIIはこの逆にMCQのＢ問題の試験、PMPではグループＩとは別の２問と共通の１問を受け、ペーパー試験ではＡ問題、およびPMPで受けなかった２問に対応するペーパー試験10題を受験した。この間Ａ、Ｂの受験生は別室でコンピュータ、あるいはペーパー試験を受けるか、あるいは同時に試験を行い、両グループの受験生が試験内容を相互に話し合える時間はないようにした。

３　総受験者数

慈恵医大、日大、東京女子医大、埼玉医大、自治医大、獨協医大の５年生、（２大学では６年生を含む)を対象に、男性264名、女性216名、計580名を同一施設、同一学年の学生を２群に分けて試験を行った。

（倫理面への配慮）

大学名、学生の氏名は用いず、コード番号で表示した。

Ｃ．研究結果

１　MCQにおける検討

MCQ試験による得点を表１および図１にまとめた。平均点では６年生に行なったＣ大学と、Ｆ大学の６年生が60点台であったのに対し、もっとも低い大学は30点台と大学群間で乖離した数字となった。コンピュータ試験(COM)とペーパーテスト（MAR）の結果において、２群間には強い相関がみられ、平均点でCOMの方が0.24低い結果が得られた。男性でも女性でもともに同様の結果であり、各大学でみてもCOMの平均点の方が低かったが、６年生に行なったＦ大学の結果のみが、COMの平均点の方が高かった(ここでは後に述べる４問の問題を除外した数値で示してある)。

図１をみるとコンピュータ試験（COM）に比してペーパー試験（MAR）でとくに得点の低い群と高い群が少数であるが存在する。ｒ＝0.70でほぼよい相関といえる。男女別にみると、図でみると男性の方がより高得点にみえるが、これは５年生に女子学生だけの大学が含まれていたためであると考える。

次に得点分布を偏差値で比較すると、図４に示すように、Ａ問題、Ｂ問題ともにほとんど二つのグラフは重なる（CBT:コンピュータ試験、MAR:ペーパ

一試験）。

個々の問題についての得点差について検討したのが図５である。ほとんどの問題でコンピュータとペーパーテストの間に正答率に差はないといえる。このうちＡ８、Ａ50、Ｂ27、Ｂ44は紙とコンピュータ問題で問題･解答内容に違いがあり除外することにした。Ａ７は選択問題Ｘ２であった。したがって他のデータ、図表ではこれらを除外した結果を示してある。Ｂ10のみは両者の正答率に大きな差がでているが、その意味を説明できなかった。４問を除外した結果、平均点±標準偏差はコンピュータテスト 50.88±15.01 に対してペーパーテストは 51.12±14.50 となり、平均点ではほぼ同じになった。

図６は個々の得点を２群に分けて比較したものである。平均点、最高点、最低点などはほぼ同じであり、両群間に有意差はない。

２．ＰＭＰ試験（PMP）と従来の紙試験（MAR）との比較

PMP の問題５題のうち３題を施行しているので、全員の３題の平均得点と、全員のそれに対応する MCQ10 問の正解率を示す（表２）。

１）PMP 試験と対応するマークシート試験との比較

３題の PMP 問題の得点と MCQ の得点との間には強い相関はなかった（相関係数 r=0.439 ）。男女比で見ると女性にやや相関がみられた（図７）。これに対して MCQ の得点と PMP に対応するペーパーテストの 10 問の得点で比較するとこれでは弱い正の相関が見られる。いわゆるペーパーテストでは全体の MCQ と同様に知識を問う試験問題になっていることから、PMP と異なった相関を示すものと考えられる（図８）。また PMP のコンピュータ問題とペーパー試験の直接の対比をみると、この散布図（図９）に示すように、両者間には相関がなかった。

２）PMP 問題での検討

問題のテーマごとにみるとそれぞれ別の傾向がある。アニサキスでは解答項目が３項目で得点もその３段階に分かれる。特に中間層は少なく、全然できないか、全て答えたか、どちらかの者が多い。一方喘息は問題の内容と解答枝の項目（治療法）が多種であり、得点分布もこれによって多岐に分かれている。肺梗塞も治療法が多くあるために得点分布も広がるが、途中に大きなギャップがある。イレウスも能力のあるものとないものの差が明確になっており、かなり能力のあるグループを入れると三群に分かれる（図 10、11、12, 13）。

３．運用上の問題

途中で作動しないことが埼玉医大で発生した。またインストールにおいて、MCQ で 82:381 PMP で 116:402 とＬＡＮ上のクライアントに配信する方法では、各施設の設定に影響されて多少の問題が起こった。画像の情報がとくに階段教室などでは周囲の学生に見えてしまい、これがヒントになることが考えられた（図１４）。

Ｄ．考察

今回用いた問題が問題として最善のものだけとは言えないので、明確な結論を出すことはできないが、これまで行われていなかった紙の試験とコンピュータの試験を直接客観的に比較する良い機会であった。

MCQ 問題における COM と MAR の間には大きな差は見られなかった。個々の問題について、正答率に特徴的な差はなかった。形式によってコンピュータとの比較において差異はなかった。男性の方がやや CBT で得点が高い傾向が伺えた。アンケートでは画像などが見やすいとの意見があったのに対して、メモをとりたい、目が疲れるなどの意見が出ており、操作性については大きな問題として上げる意見はなかったが、直接の両者を比較した意見ではペーパー試験の方がよいという意見の方が多かった。紙試験に慣れている受験生の意見として当然のことと思われる。

PMP での得点と全体の MCQ の得点と比較することで、PMP の独自性と蓋然

性を検討した。正の相関は得られたが、とくにPMPでは低い得点として存在する群が認められ、より明確に能力の低い学生を検出するものか、統合的意思決定能力に欠けるもの、あるいはコンピュータを苦手とするものを検出しているものと思われた。一方でMCQの結果とある程度の正の相関がみられることから、全く異なった能力を評価しているわけではないと考える。

PMP問題と対応するMCQの問題は必ずしも内容が一致しておらず、相互の比較をすることはできないと考えられた。ひとつのPMPの中で治療や診断の解答数の多いほど、得点分布がひろがり、喘息などをみても、かなり均一に広がっている。これに対して有効回答が3つしかないアニサキスはとくに極端な得点分布になっている。また全く回答できない例も多く見られ、ペーパー試験に比べ、得点の低いものの検出力に優れているともいえる。一方で診断名、治療数など解答数により得点分布が大きく異なることから、得点で評価するためにはある程度の解答要求数があったほうがよいのかもしれない。しかし逆に何を持って評価するのかがあいまいになることを考えると、必ずしも細かいほうがよいのでなく、アニサキスの問題のように3つくらいのカテゴリーに分けるほうが正確な指標であって、これがPMPの限界なのかもしれない。問診の中で適切なものを選んだかどうかの評価まで加えることでより正しい評価を示すものであるといえるが、従来そこまで分析したシステムはなかったし、問題を作るのが大変であろう。しかしせっかくの総合能力の試験であるとするなら問診、所見、検査、診断、治療の総合点で評価すべきものかもしれない。問診なども、たとえば決定的なポイントになるものだけにつけるなどの工夫をすればよいかもしれない。

アンケート結果ではMCQよりも、PMPの方が好ましい結果も得られているが、操作性の問題で批判がある。特に検索方式であるが、これらは改善の余地があると考えられる。

運用上の問題としてインストールの方法のトラブルであったと思うが、これだけの数を正確に同時に動かすことが求められるとなると、試験のリスクが大きい。端末の故障が起こる確率から100台あたり数台以上の予備機の準備は必要である。

PMPは画像の含まれる問題ではこれが周囲に見える状況になると大きなヒントを与えることになりかねない。このためには隔離した環境、あるいは広いスペースが必要になる。

E．結論

MCQ問題をコンピュータ試験で行なうことに大きな問題はなく、またその得点は紙のテストの結果とほぼ一致した。しかしながらメモ機能などを希望する声があり、また同時に多数の受験者を扱うことの問題は残っている。PMP問題の結果はMCQの結果と強い相関を示さず、PMPでは統合能力、意思決定能力など別の能力を見ていると考えられた。またとくに能力の低いものを明瞭に判別しえた。

現在ペーパーテストをコンピュータ化することの緊急性はないが、今後多くの試験がコンピュータ化していくことは容易に想像される。メモ機能などいくつかの技術的な改良とともに、大量の問題作成方法、および同時に実施する方式をとるのか、別の日時で試験を行う方式をとるのかの検討が必要なところにきている。これにより準備は大きく異なる。

あるいは医師免許の更新制が開始されるのであれば、このようなところから開始するという方法もあるかもしれない。

F．健康危険情報

なし

G．研究発表
1．論文発表
なし
2．学会発表
高林克日己　細田瑳一　福島統ほか　医師国家試験へのコンピュータ試験の導入の検討　第 37 回日本医学教育学会大会（予定）

H．知的財産権の出願・登録状況（予定を含む。）
なし

(表１)　MCQ問題のコンピュータ試験とペーパーテストの結果

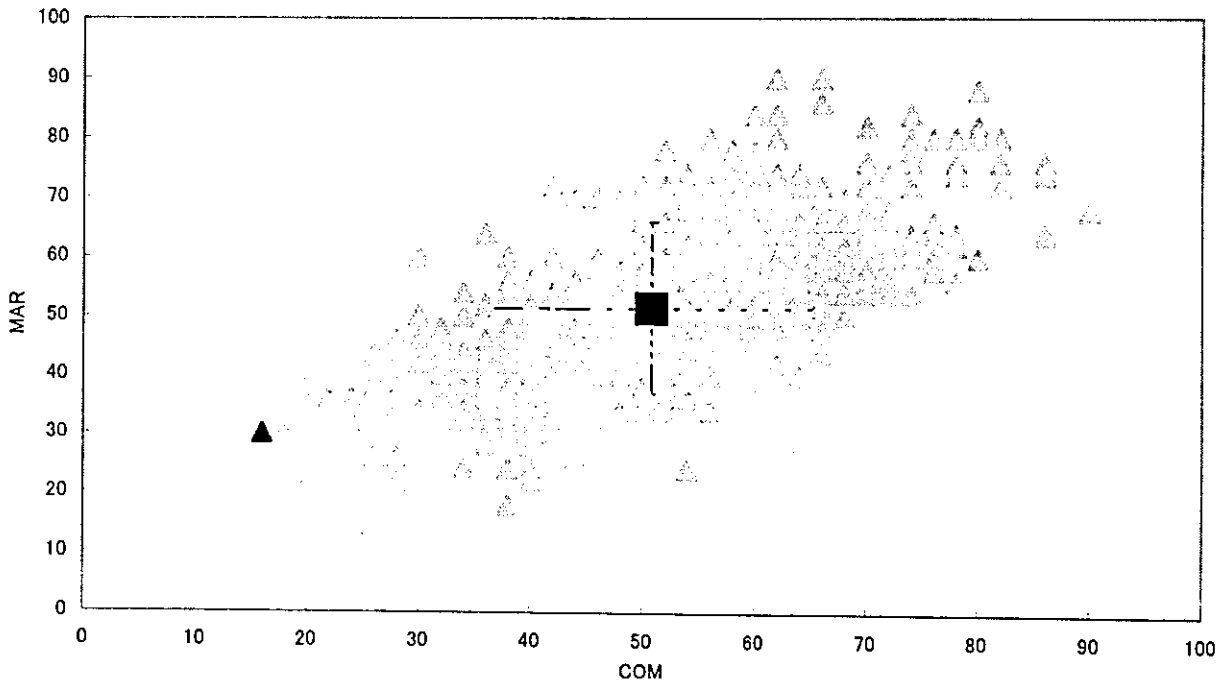
コンピュータ試験 (COM)		ペーパー試験 (MAR)	
全体		全体	
被験者数	504	被験者数	498
平均点	50.88	平均点	51.12
標準偏差	15.01	標準偏差	14.50
最高点	92	最高点	90
最低点	17	最低点	18
A大			
被験者数	23	被験者数	23
平均点	44.20	平均点	44.75
標準偏差	9.01	標準偏差	7.86
最高点	60	最高点	62
最低点	27	最低点	30
B大			
被験者数	112	被験者数	112
平均点	42.06	平均点	42.52
標準偏差	11.58	標準偏差	11.48
最高点	69	最高点	78
最低点	17	最低点	22
C大			
被験者数	104	被験者数	104
平均点	61.78	平均点	62.23
標準偏差	12.42	標準偏差	12.45
最高点	90	最高点	90
最低点	21	最低点	24
D大			
被験者数	89	被験者数	86
平均点	54.99	平均点	54.30
標準偏差	11.19	標準偏差	9.64
最高点	79	最高点	76
最低点	29	最低点	32
E大			
被験者数	92	被験者数	90
平均点	37.86	平均点	39.58
標準偏差	8.85	標準偏差	8.84
最高点	67	最高点	64
最低点	19	最低点	18
F大(５年)			
被験者数	26	被験者数	24

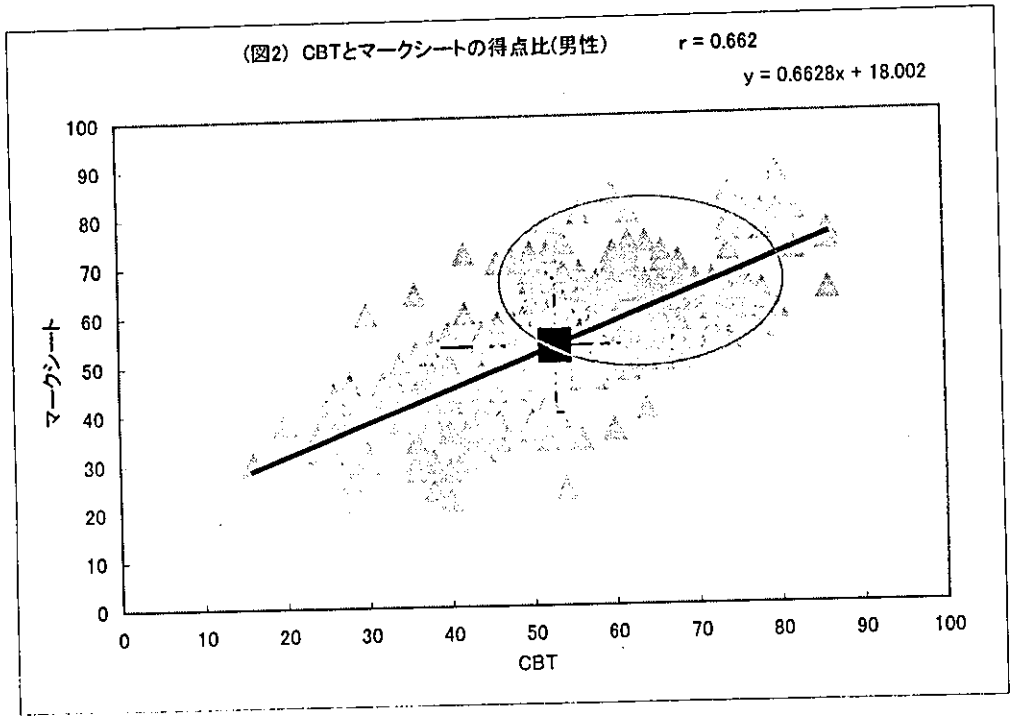
平均点	44.87	平均点	45.58
標準偏差	8.96	標準偏差	10.37
最高点	63	最高点	70
最低点	25	最低点	30

F大(6年)

被験者数	59	被験者数	58
平均点	66.74	平均点	65.90
標準偏差	11.31	標準偏差	10.87
最高点	92	最高点	88
最低点	40	最低点	36

(図1) MCQにおけるCOMとMARの得点比
R=0.70




(図2) CBTとマークシートの得点比(男性)
r = 0.662
y = 0.6628x + 18.002
マークシート
CBT
0
10
20
30
40
50
60
70
80
90
100

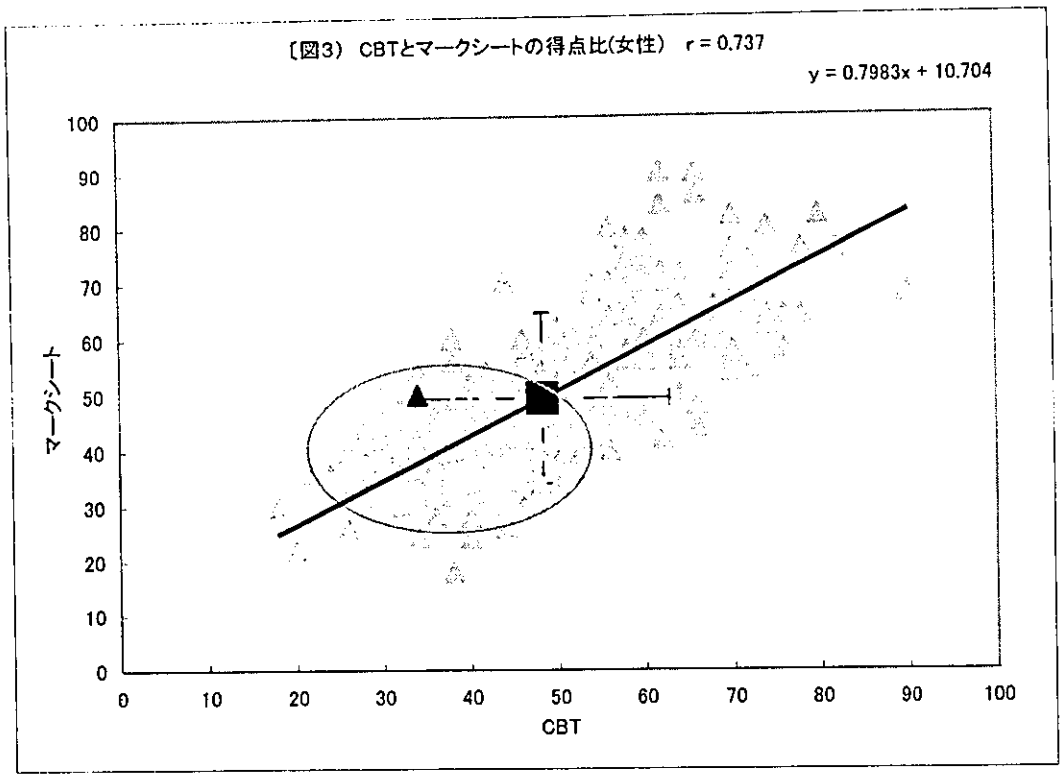
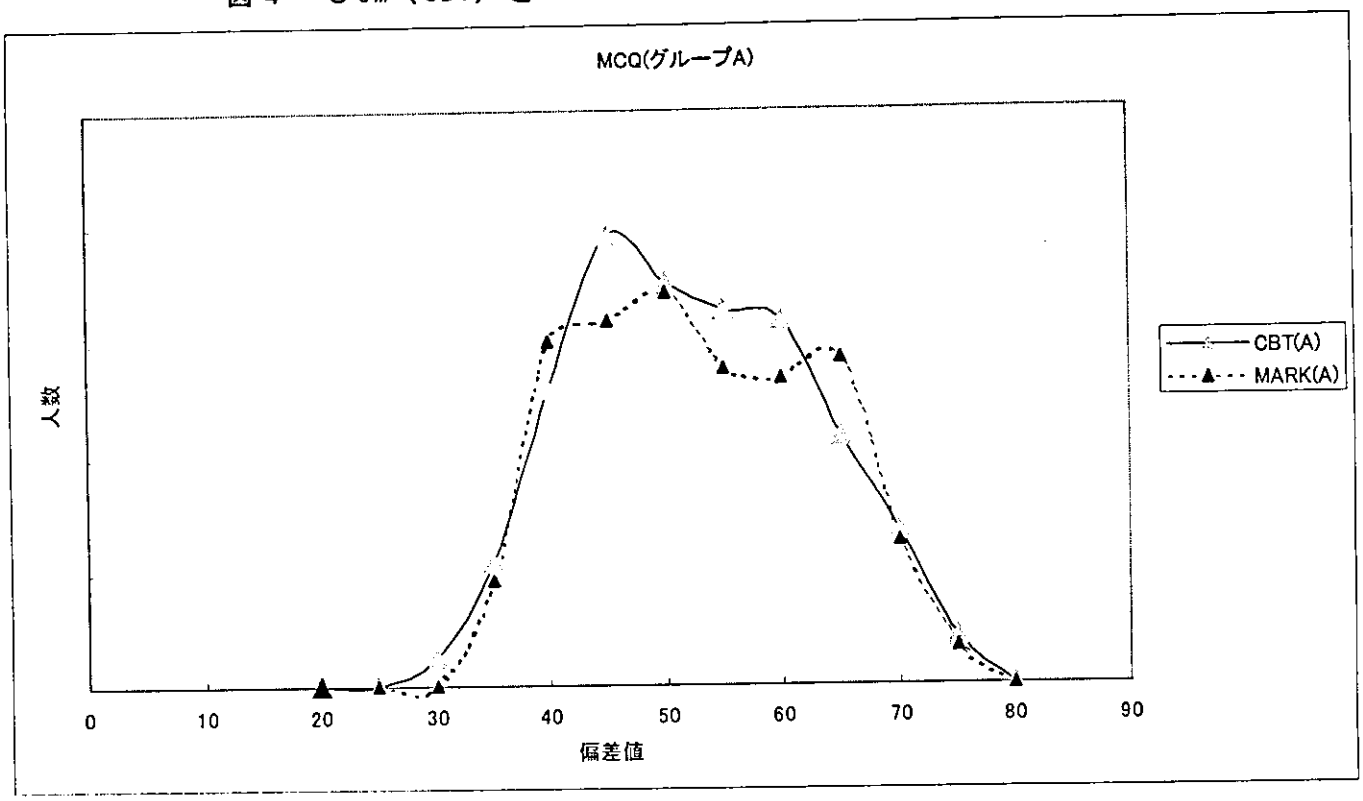
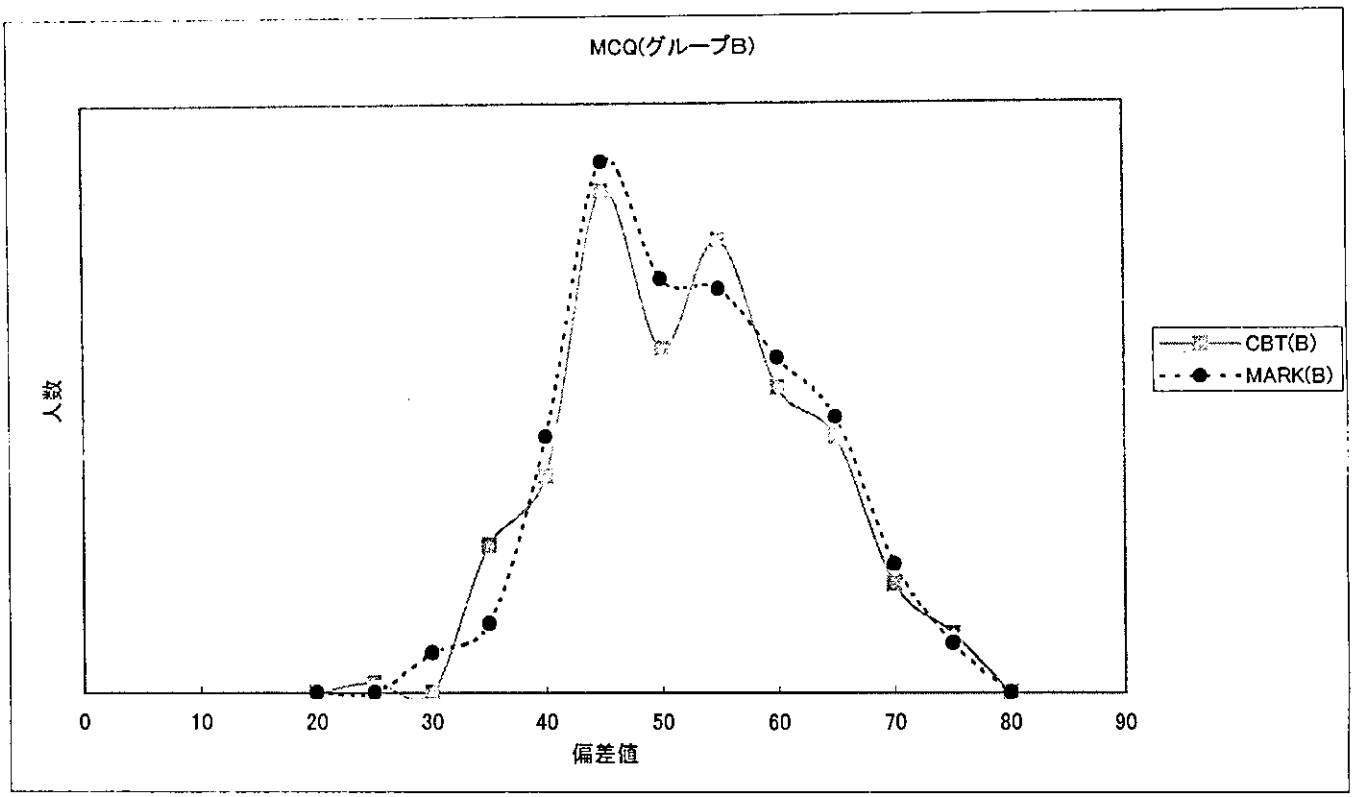
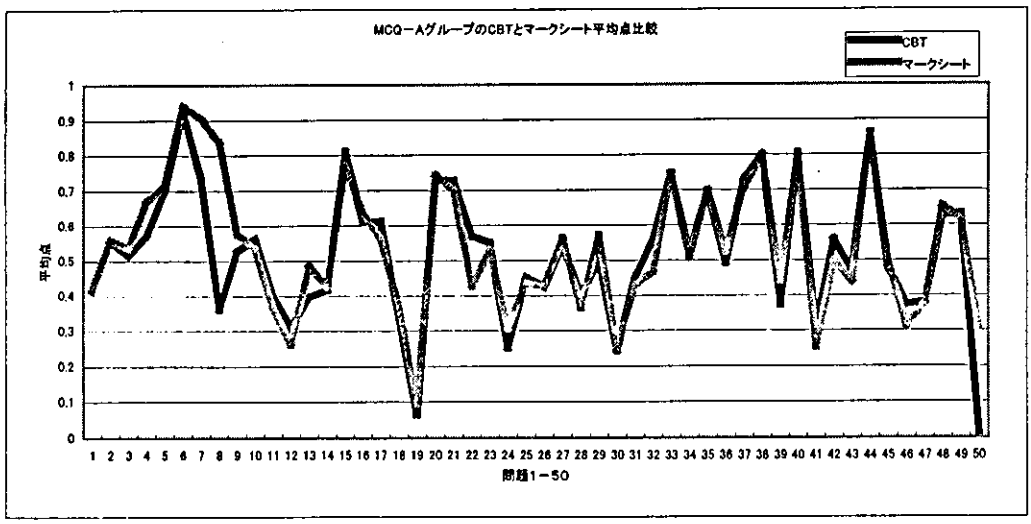

〔図3〕 CBTとマークシートの得点比(女性) r = 0.737
y = 0.7983x + 10.704
マークシート
CBT
0
10
20
30
40
50
60
70
80
90
100

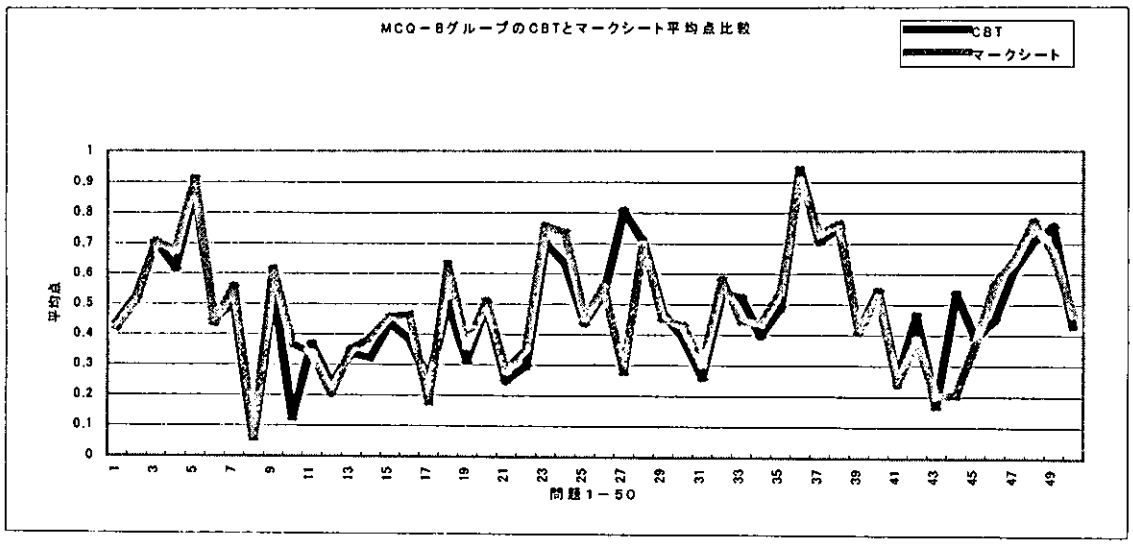
図４　ＣＯＭ（ＣＢＴ）とＭＡＲにおける各問題の正答率





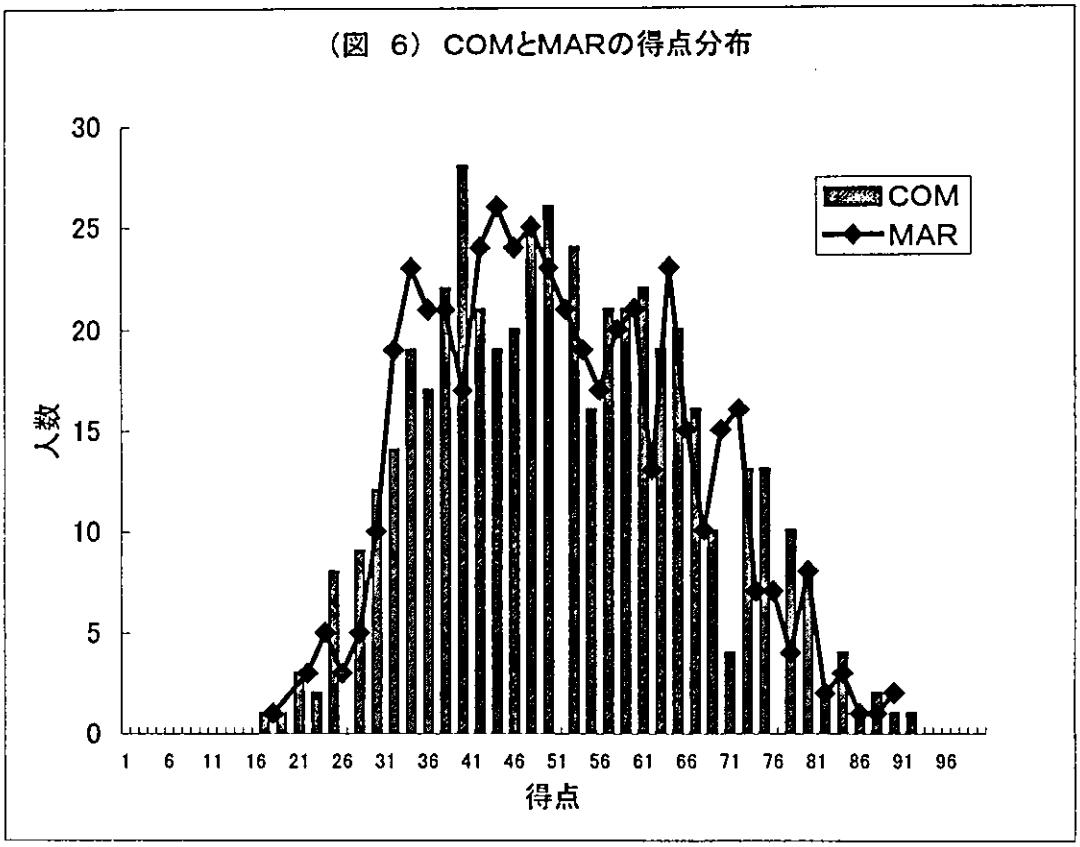

MCQ－Aグループの CBTとマークシート平均点比較
CBT
マークシート
平均点
問題1－50

（図５A）


MCQ－Bグループの CBTとマークシート平均点比較
CBT
マークシート
平均点
問題1－50

（図５B）

（図 6） COMとMARの得点分布



（表 2 ）PMP 問題の実施結果

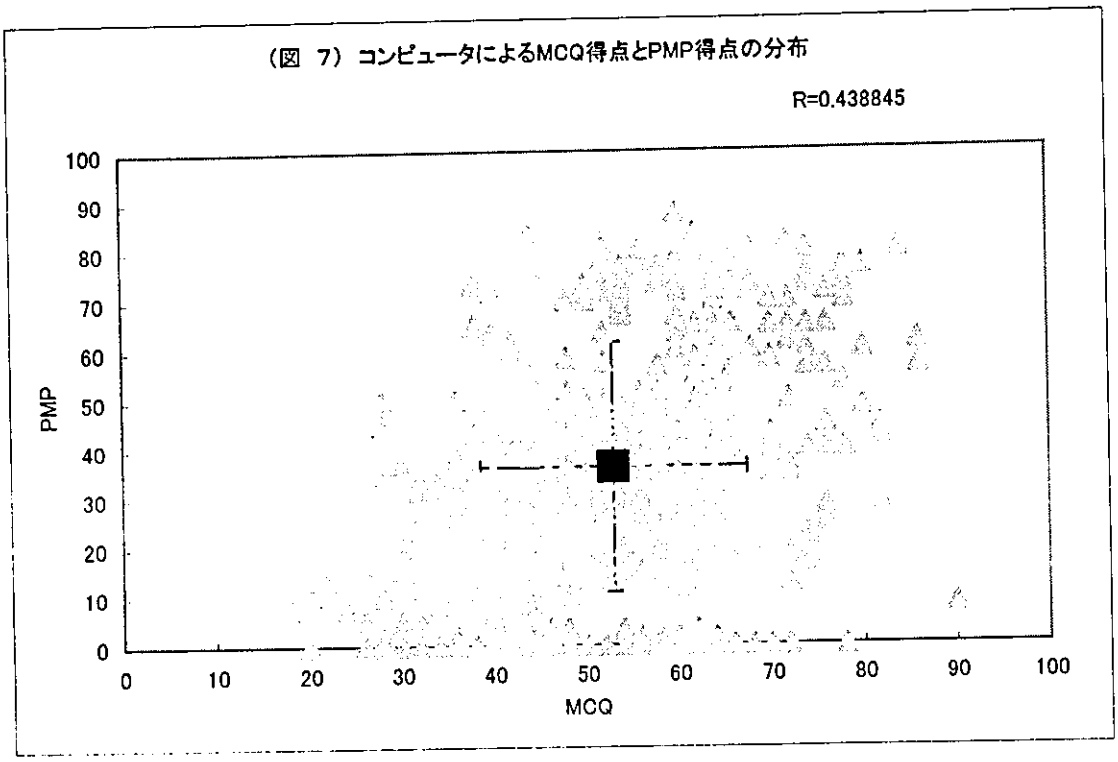
ＣＢＴ

全体		アニサキス		喘息		肺梗塞		イレウス		過換気	
被験者数	505	被験者数	244	被験者数	255	被験者数	240	被験者数	231	被験者数	485
平均点		平均点	61.27	平均点	33.82	平均点	29.40	平均点	29.87	平均点	31.8
標準偏差		標準偏差	45.43	標準偏差	27.49	標準偏差	20.74	標準偏差	31.41	標準偏差	4
最高点		最高点	100	最高点	95	最高点	85	最高点	100	最高点	37.5
最低点		最低点	0	最低点	0	最低点	0	最低点	0	最低点	100
											0

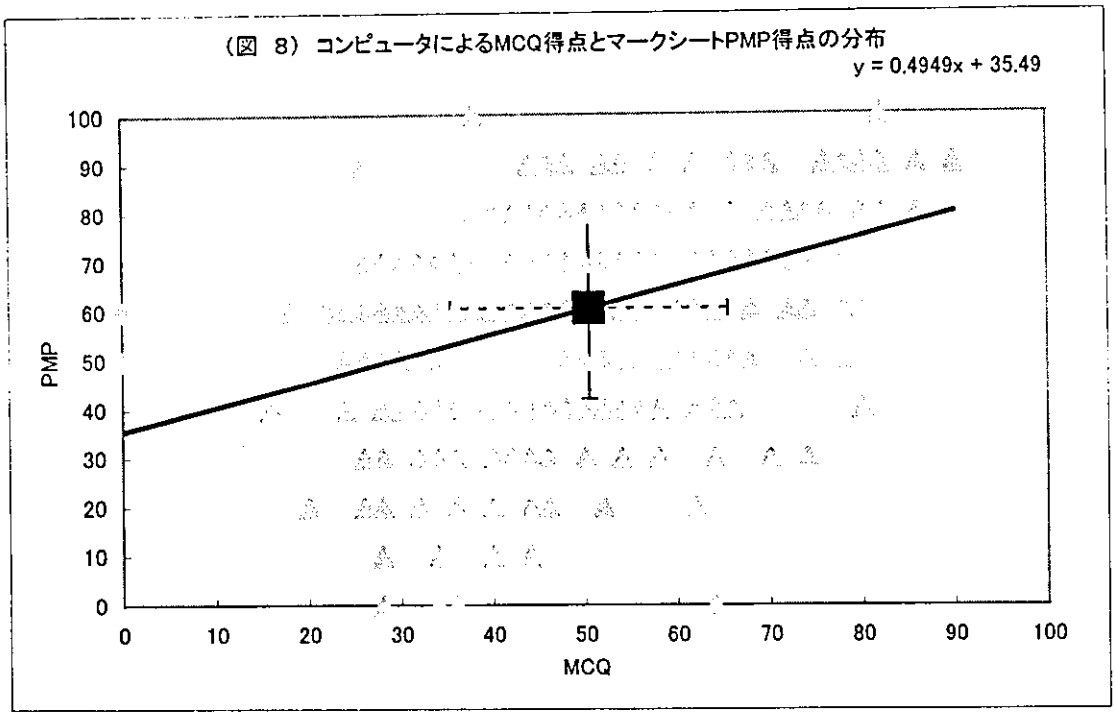
MAR

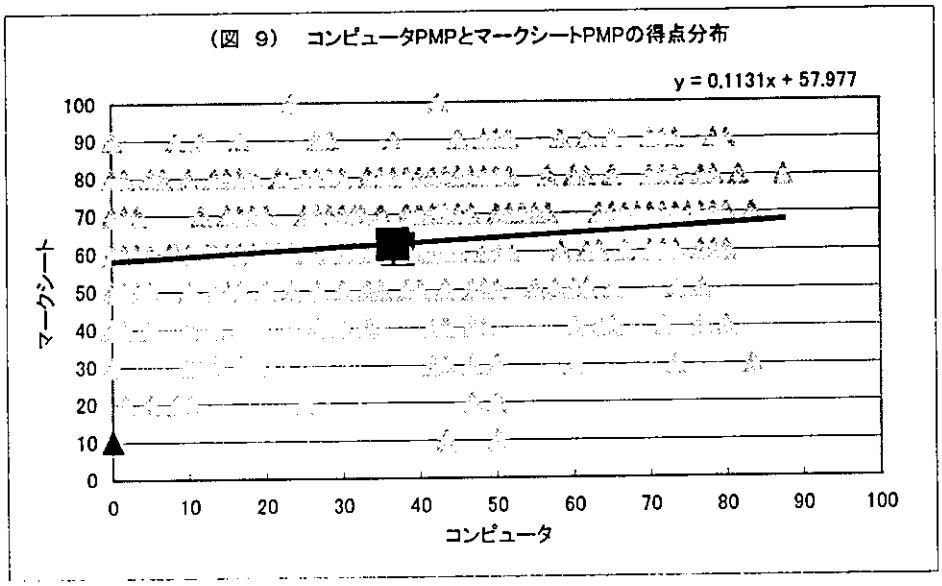
全体		アニサキス		喘息		肺梗塞		イレウス	
被験者数		被験者数	254	被験者数	244	被験者数	254	被験者数	244
平均点		平均点	60.39	平均点	51.39	平均点	74.96	平均点	54.10
標準偏差		標準偏差	19.76	標準偏差	22.14	標準偏差	22.42	標準偏差	20.88
最高点		最高点	100	最高点	100	最高点	100	最高点	100
最低点		最低点	0	最低点	0	最低点	0	最低点	0

（図 7）コンピュータによるMCQ得点とPMP得点の分布

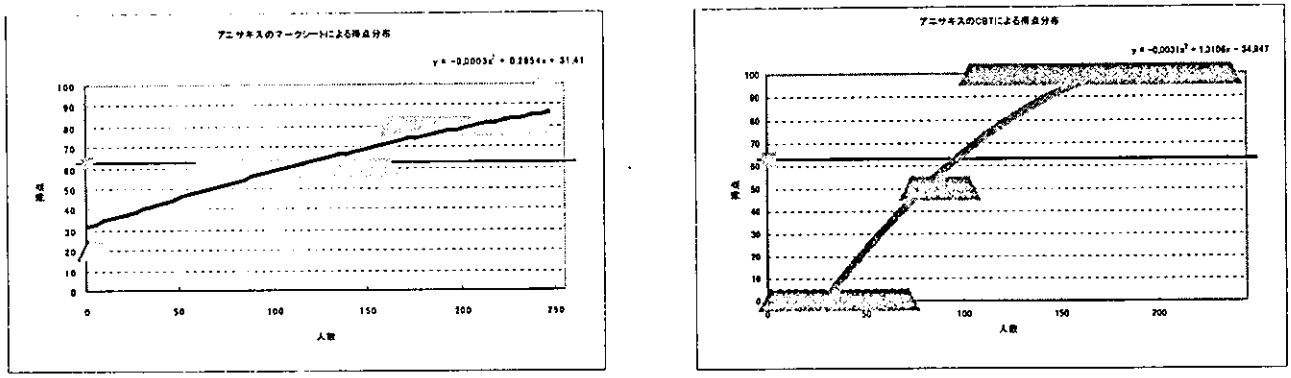


（図 8）コンピュータによるMCQ得点とマークシートPMP得点の分布

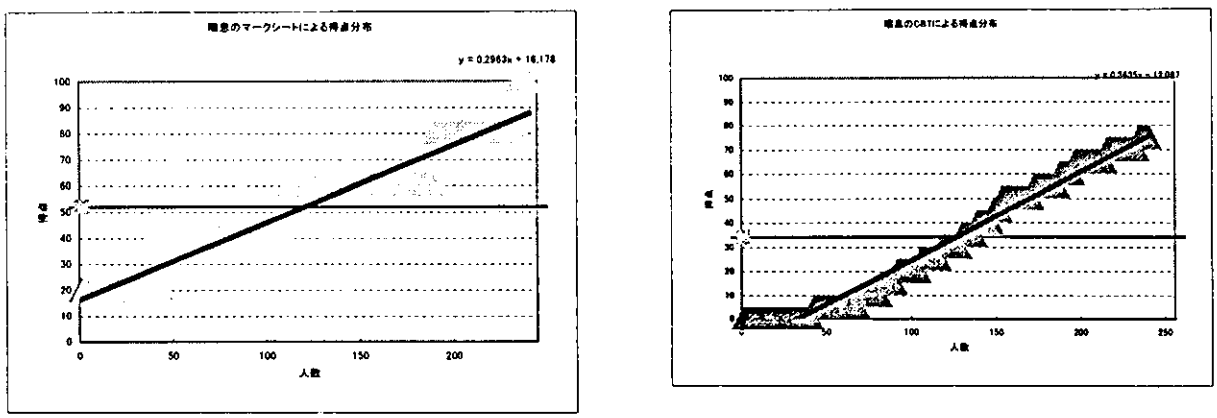




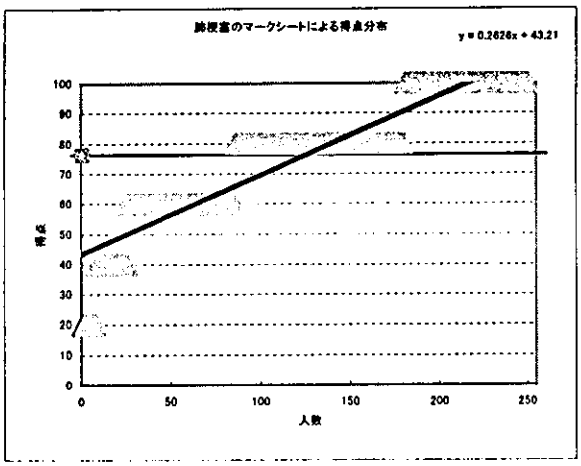
（図　１０）　アニサキスの問題の得点比較（左がマークシート、右が PMP）

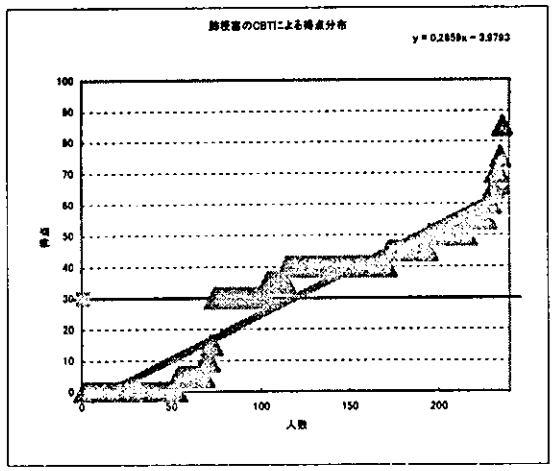


（図　１１）　気管支喘息の問題の得点比較（左がマークシート、右が PMP）

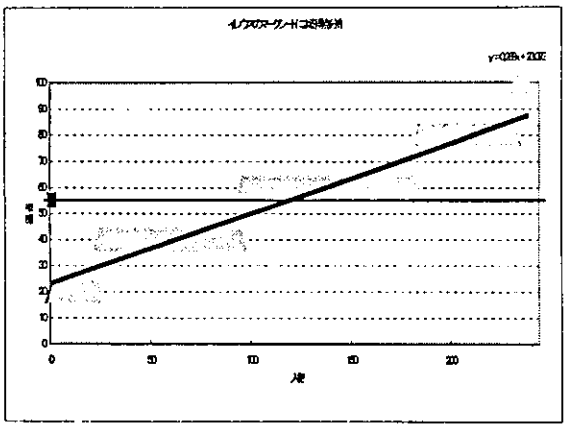


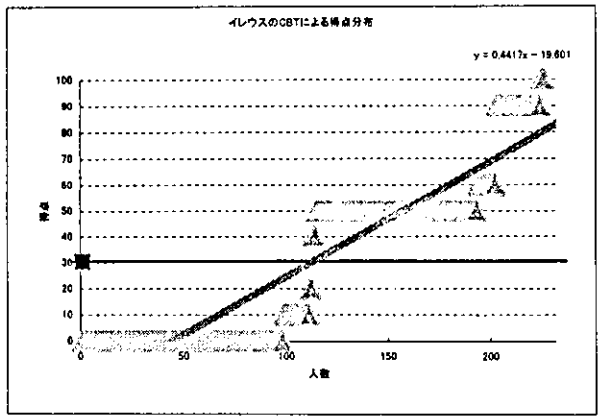
（図　１２）　肺梗塞の問題の得点比較（左がマークシート、右が PMP）

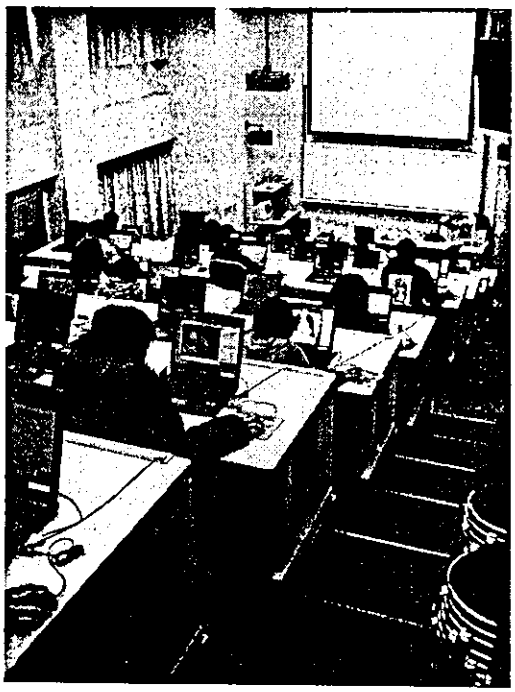




（図　１３）　イレウスの問題の得点比較（左がマークシート、右が PMP）







（図　１４）階段教室の試験風景

厚生労働科学研究費補助金（医療技術評価総合研究事業）

分担研究報告書

試験問題のコンピュータ化と諸外国の実情視察

分担研究者　高林克日己　千葉大学医学部医療情報部助教授

研究要旨

研究班で医師国家試験の模擬問題をA，B６０題づつ作成し、それぞれ筆記試験またはコンピュータ使用の試験として学生に実施し、その結果をまとめて比較し、コンピュータ化の実施可能性と利点、欠点について検討した。

A．研究目的

医師国家試験にコンピュータを利用した試験を導入する妥当性を検討する。

B．研究方法

①福島統分担研究者が作成した 50 問の試験問題２セット、計 100 問をコンピュータ化して、画像などとセットにし、ウインドウズマシンにインストール可能とした。同様に一昨年作成したツールを介して PMP 問題を作成し、これら MCQ, PMP とアンケートをまとめてインストールできるように設計した。試験中には受験者の解答のほか、試験時間、とくに PMP では各セクションに関する解答時間を収集し、これらをまとめて、ハードデイスク上に一意の名前のファイル作成するようにした。受験者にはデータベースや画像は受験中に侵入できたり閲覧できないように設計した。実際のファイルの展開を示す。図１は MCQ の問題の１例である。画像はクリックすることで大きくより鮮明な画像を表示できる。どの問題を終了したかは一覧でみることができまたどれでも番号を選ぶことでその問題に移ることができる。１選択、でも複数選択でも入力が可能である。
また試験終了時にはすぐその場で評価点が示される。
PMP 問題ははじめのオープニングシーンの情報〔図３〕以降は全て受験者が選択するもので、その結果は「カルテ」に貯まっていく〔図４〕。必要に応じて画像が展開することもある。最終的には診断と治療が行なわれ、これらの点数で評価が行なわれる。終了時には選択結果と各セクションの所要時間が表示される〔図５〕。これらを5問作成した。アンケートもこれと同様にコンピュータから行なった
②受験者のアンケートをとって情報を収集する。学生アンケートは全員の学生に試験終了時にコンピュータへの入力として行い、複数回答を許した。

C．結果

①総括に付す
②１）　MCQ について
表１、図６に PMP 試験に対する学生の評価を示した。画像がみやすい、操作が楽という他に、紙よりよいとするものが 20 件あった。いっぽうで negative な意見としてメモが取りにくい、長文を読みにくい、難しかった、紙の試験の方がよい、目が疲れるという意見であった。
２）PMP について
PMP についてのアンケートでは面白いとするものが賛成意見のほとんどを占めるのに対し、操作性の悪さ、難しさを訴えるものが多かった（表２、図７）。

D．考察

①②総括に付す

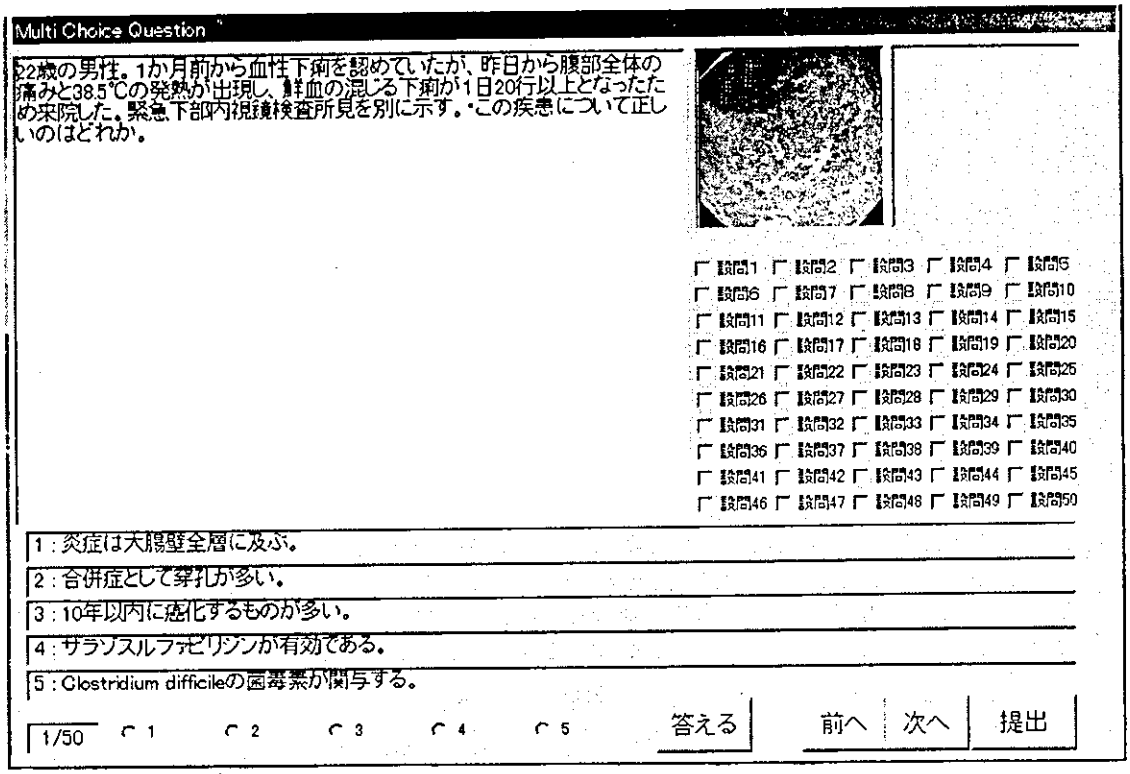


図1　MCQ の画面

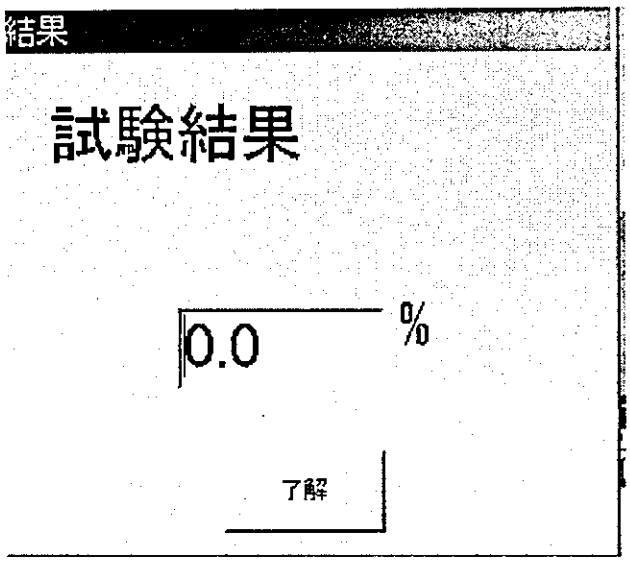


図2　試験終了時の表示

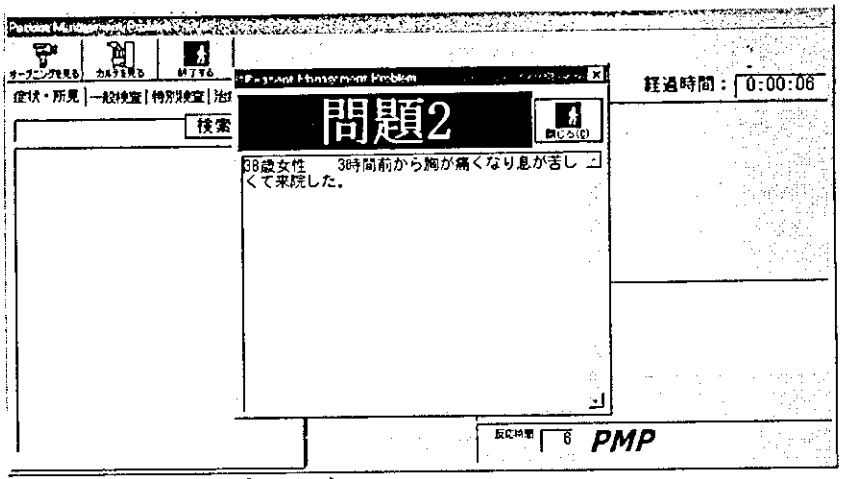


図３　PMP のオープニングシーン

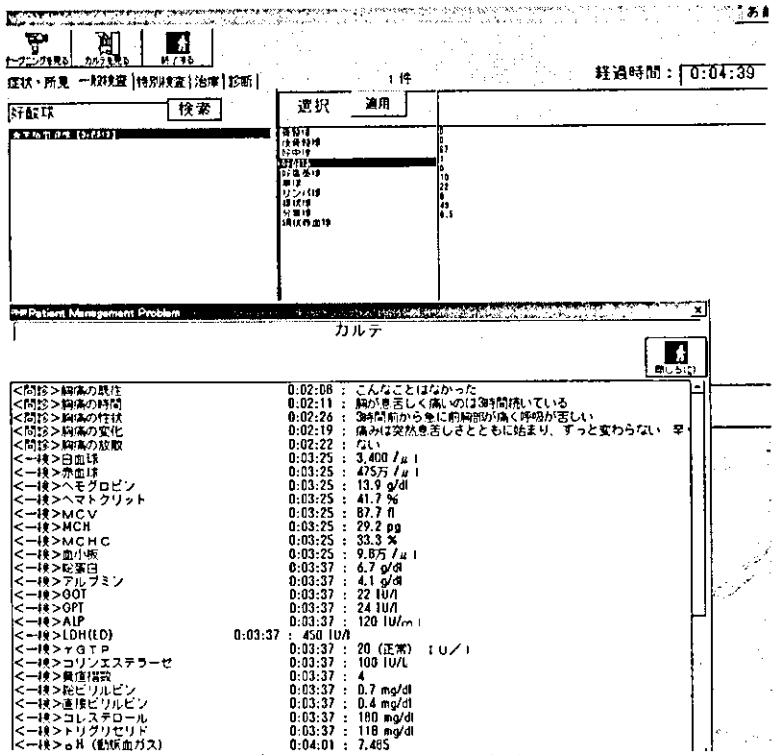


図４　PMP の項目選択の場面とカルテ内容

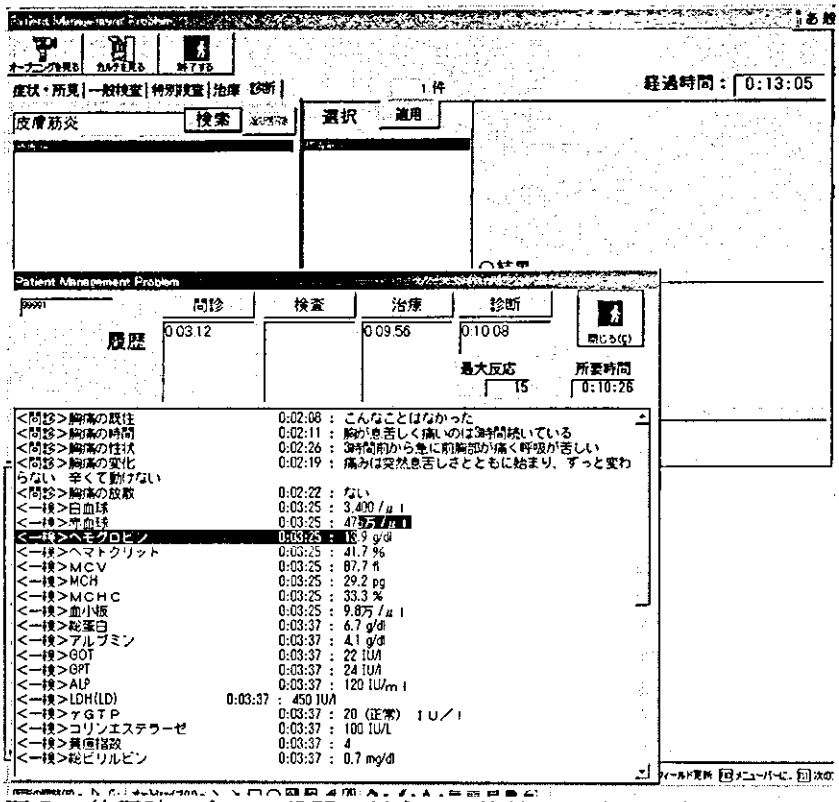


図５　終了時　全ての設問に対する回答結果と時間が表示される