医師国家試験コンピュータ化に関する研究

報　告　書

今年度は作成したコンピュータの MCQ 方式と PMP 方式の問題を受験させるとともに、同一の問題をペーパー試験で行うことで、その対比を行った。

１　対象と方法

１．１　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、外科系問題を集めた中から同一の分野でレベルの近似した対に問題になるような 100 題を抽出して、各 A,B 問題 50 題とし、それぞれにコンピュータ用とペーパー試験用とを作成した。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題を分担研究者が作成し、このうち４題については各 PMP 問題に対応するようにそれぞれに５問の MCQ 問題をペーパー試験用に作成した。

１．２　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。

１．３　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、（一大学では 6 年生も）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。

２　結果

２．１　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)

ペーパによる結果

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

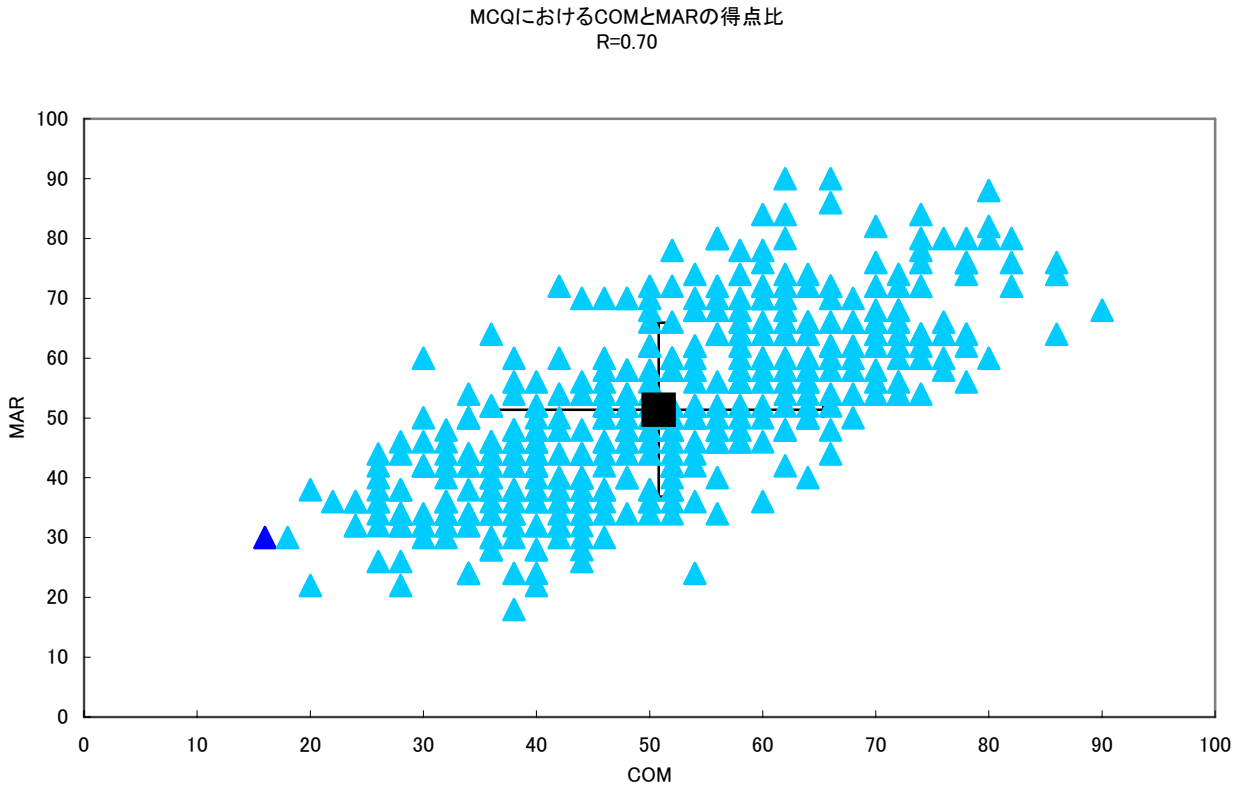
F 大(5年)

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)

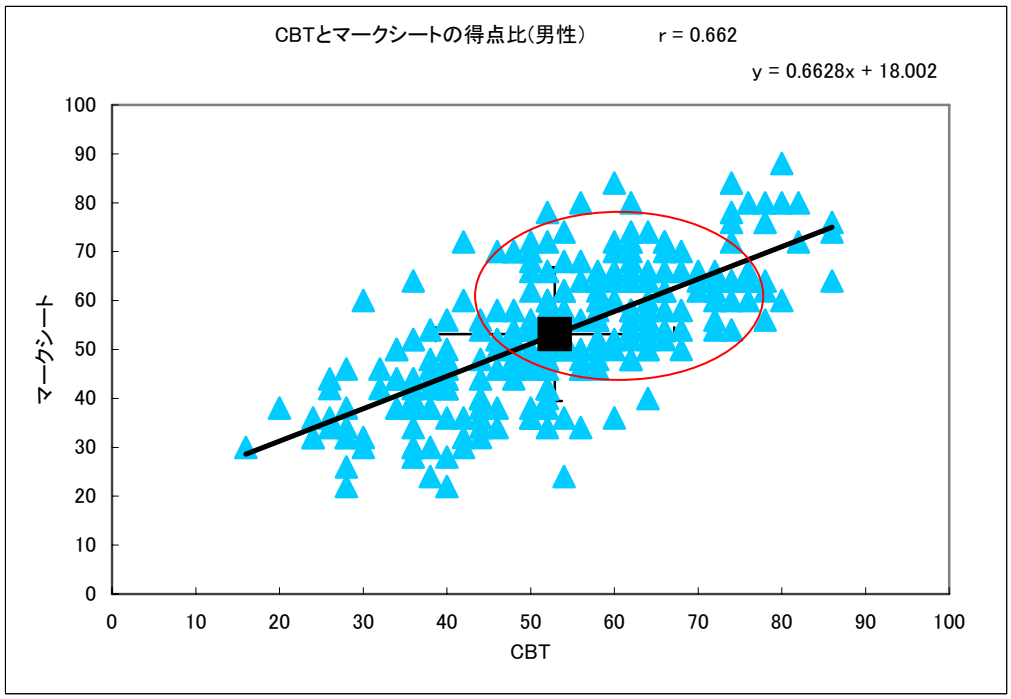
被験者数	59	被験者数	58
平均点	66.74	平均点	65.90
標準偏差	11.31	標準偏差	10.87
最高点	92	最高点	88
最低点	40	最低点	36

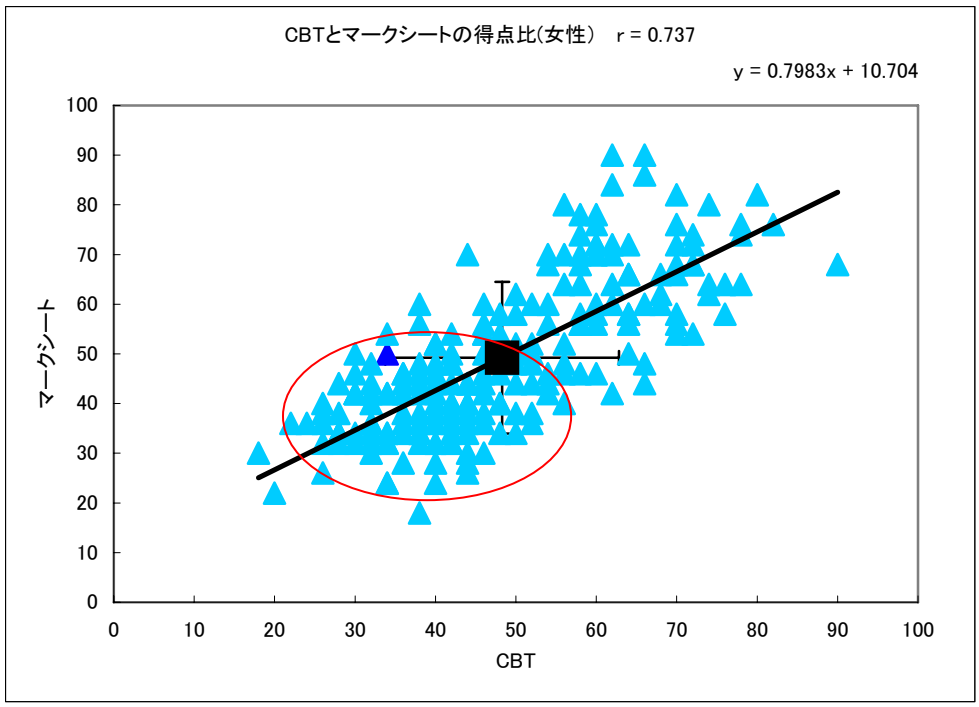
図１をみると COM でとくに得点の低い群と高い群が少数であるが存在する。 r ＝0.70 でほぼよい相関といえる。

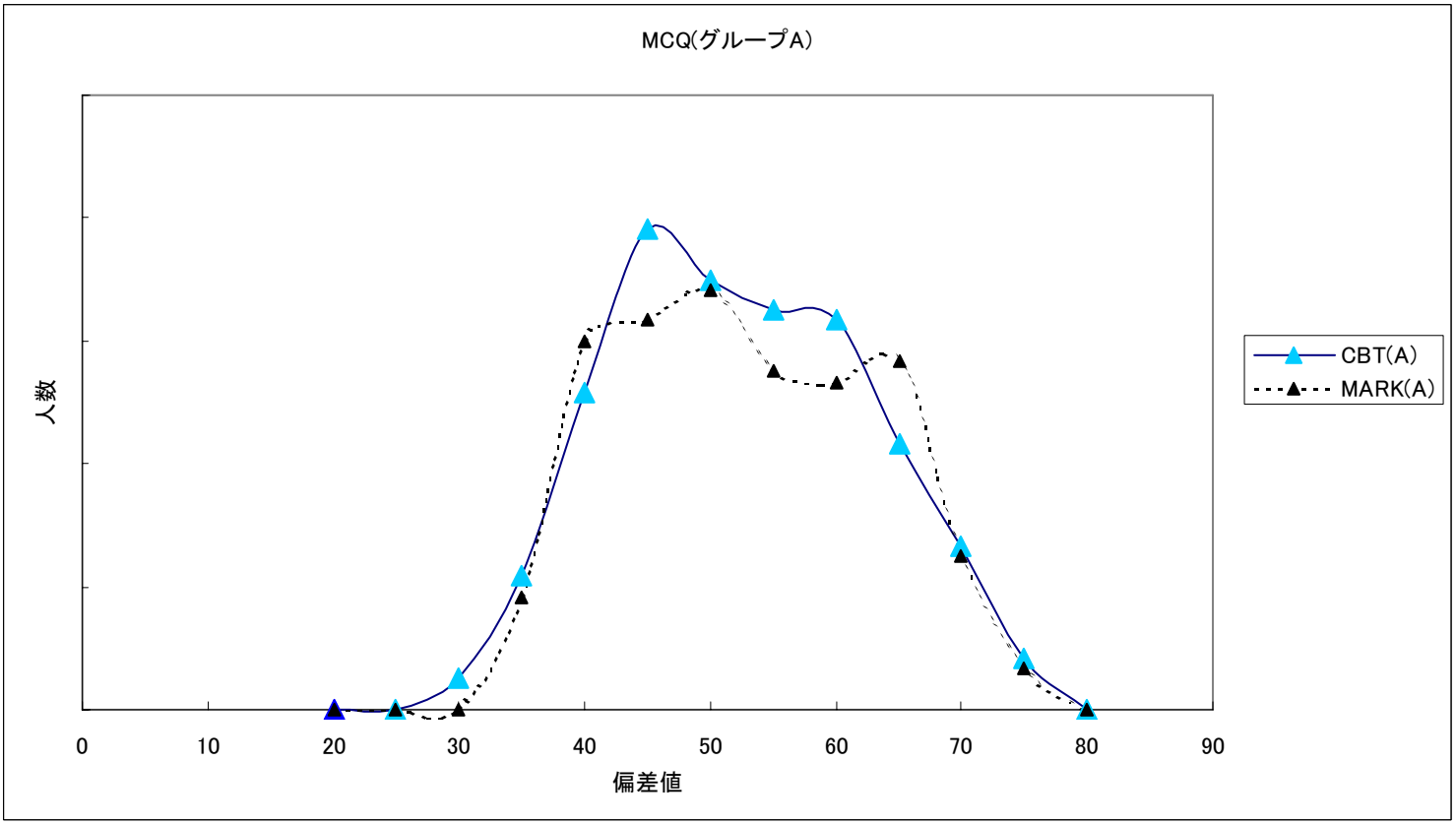


次に得点分布を偏差値で比較すると、次頁に示すように、A 問題、B 問題ともにほとんど二つのグラフは重なる（CBT：コンピュータ試験、MAR:ペーパー試験)。個々の問題についての得点差について検討したのが図３である。ほとんどの問題でコンピュータとペーパーテストの間に正答率に差はないといえる。

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 年生が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。







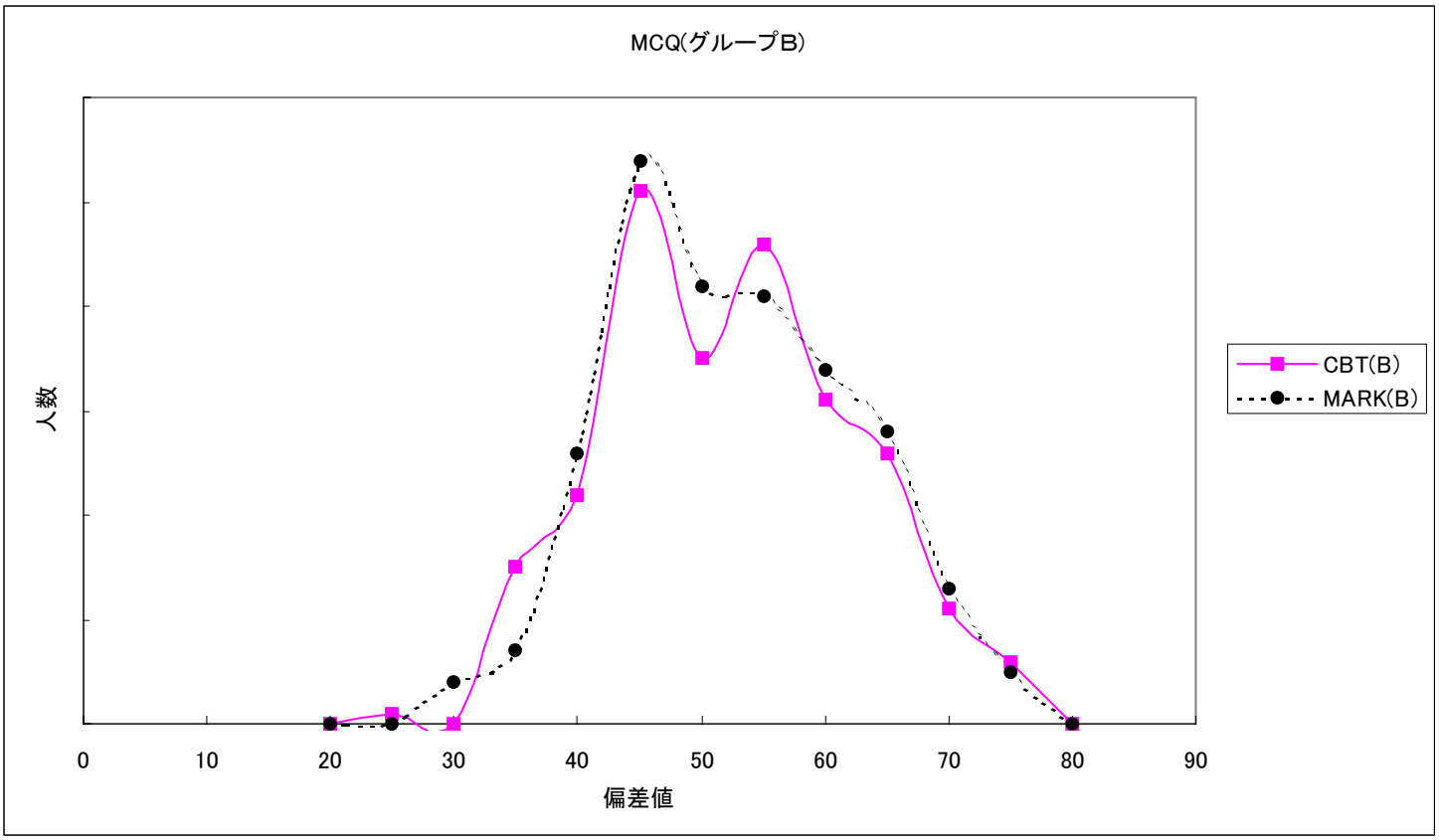
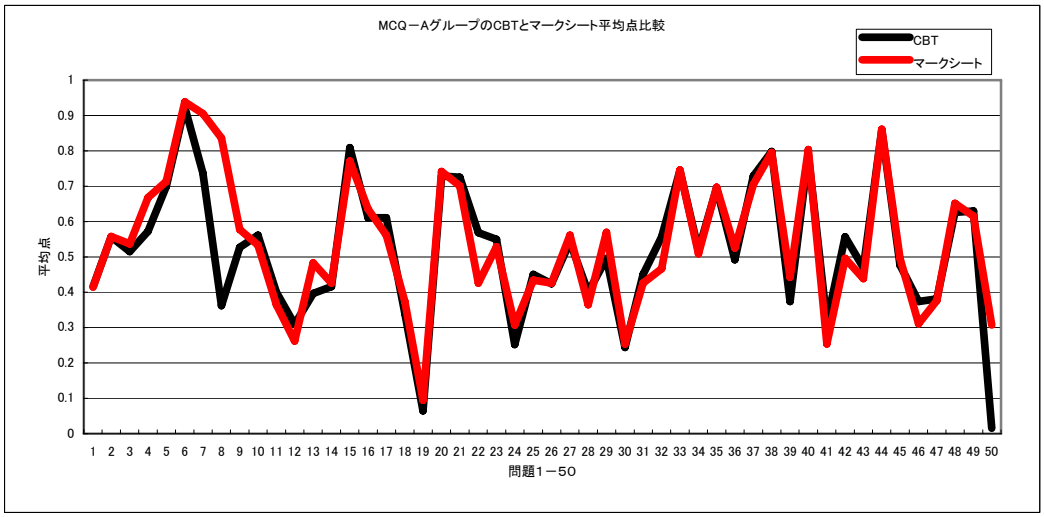
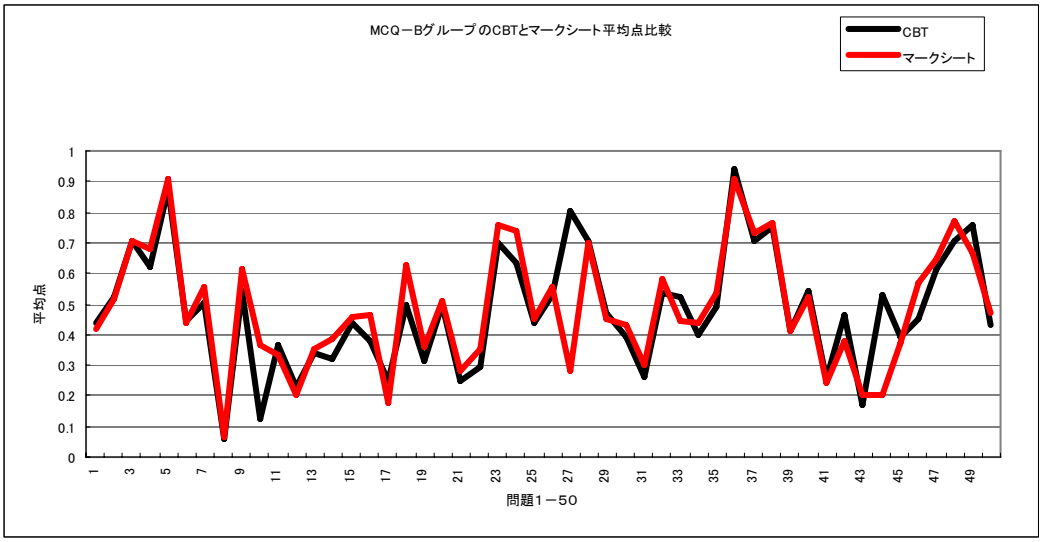
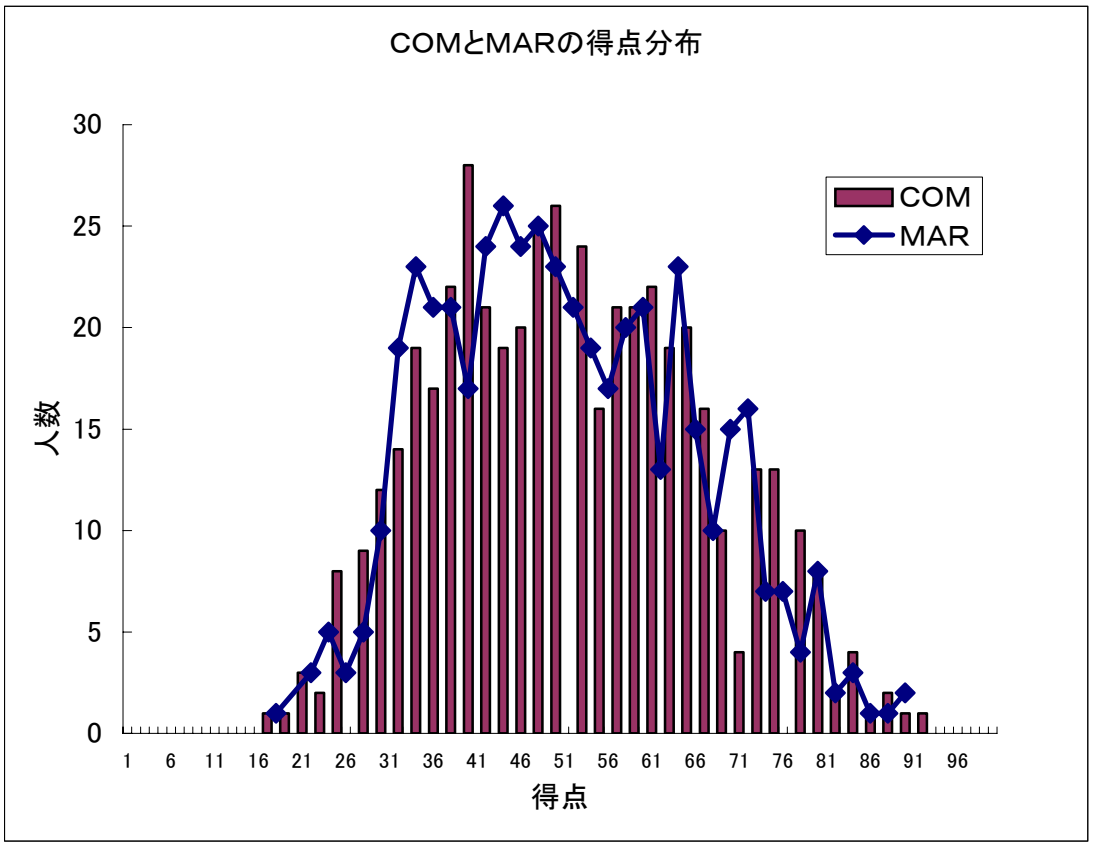


図２　ＣＯＭ（CBT）とＭＡＲにおける各問題の正答率

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 のみは差が出ているが、その意味を説明できなかった問題である。これにより平均点の差はコンピュータテスト 51.2±15.1 に対してペーパーテストは　51.4±14.5 となり、平均点ではほぼ同じになった。







上図は個々の得点を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。

４　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。

CBT

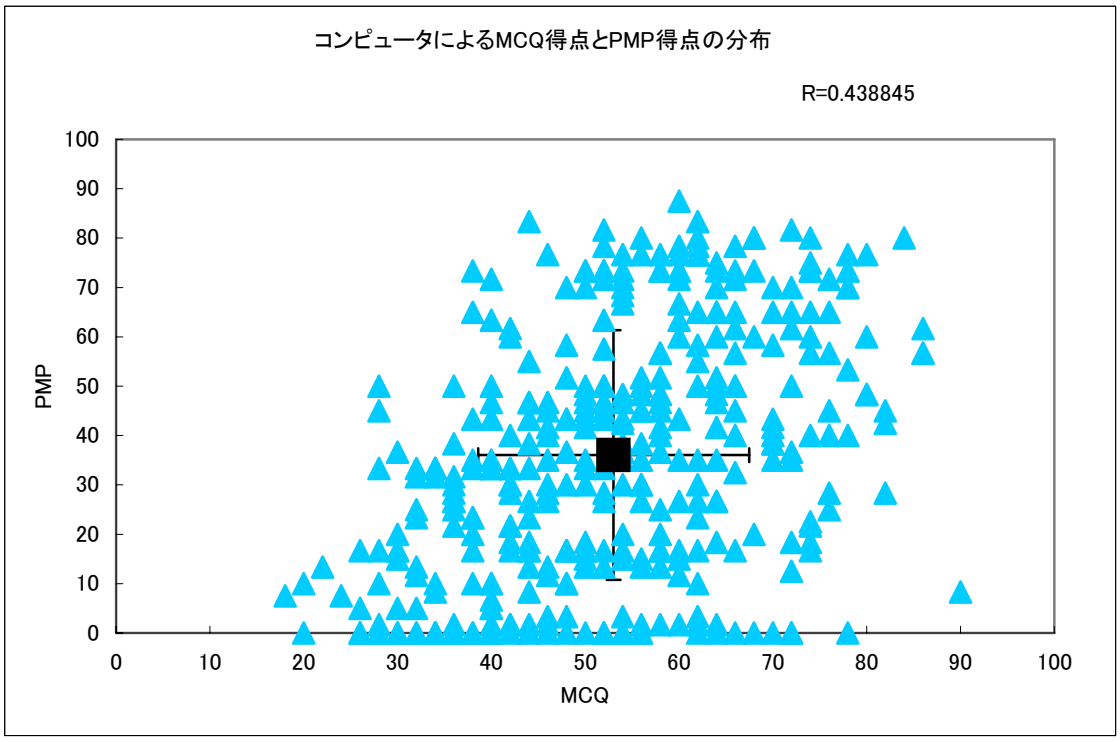
全体		アニサキス		喘息		肺梗塞		イレウス		過換気	
被験者数	505	被験者数	244	被験者数	255	被験者数	240	被験者数	231	被験者数	485
平均点		平均点	61.27	平均点	33.82	平均点	29.40	平均点	29.87	平均点	31.84
標準偏差		標準偏差	45.43	標準偏差	27.49	標準偏差	20.74	標準偏差	31.41	標準偏差	37.5
最高点		最高点	100	最高点	95	最高点	85	最高点	100	最高点	100
最低点		最低点	0	最低点	0	最低点	0	最低点	0	最低点	0

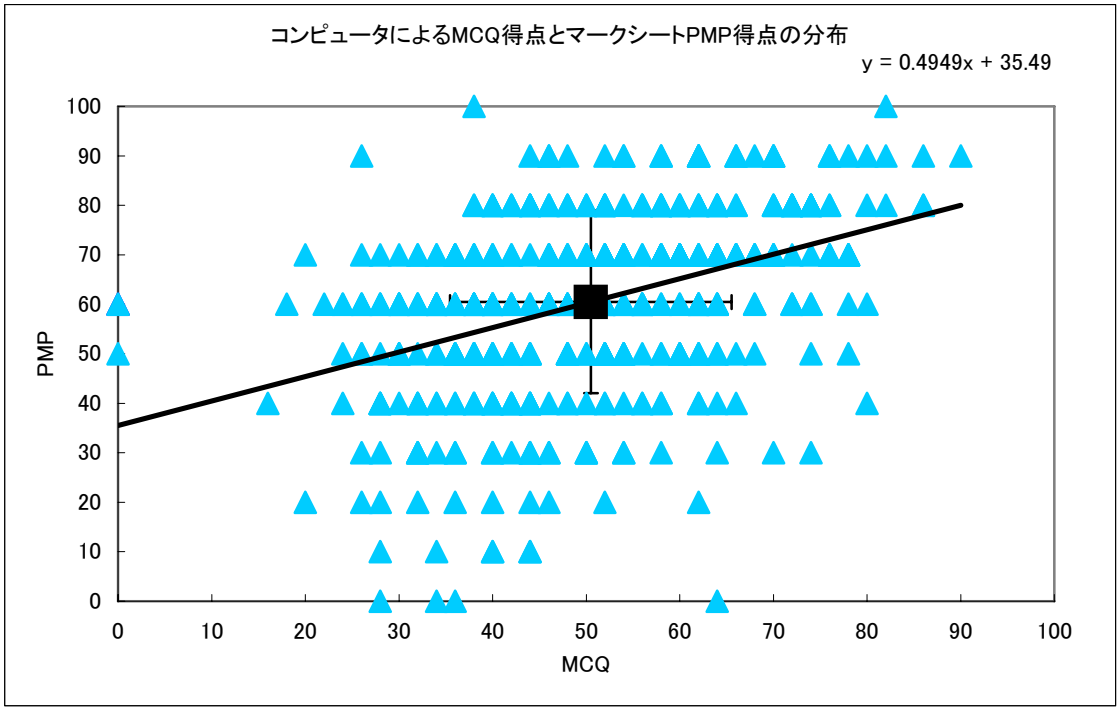
MAR

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

４．１　ＰＭＰ試験と対応するマークシート試験との比較

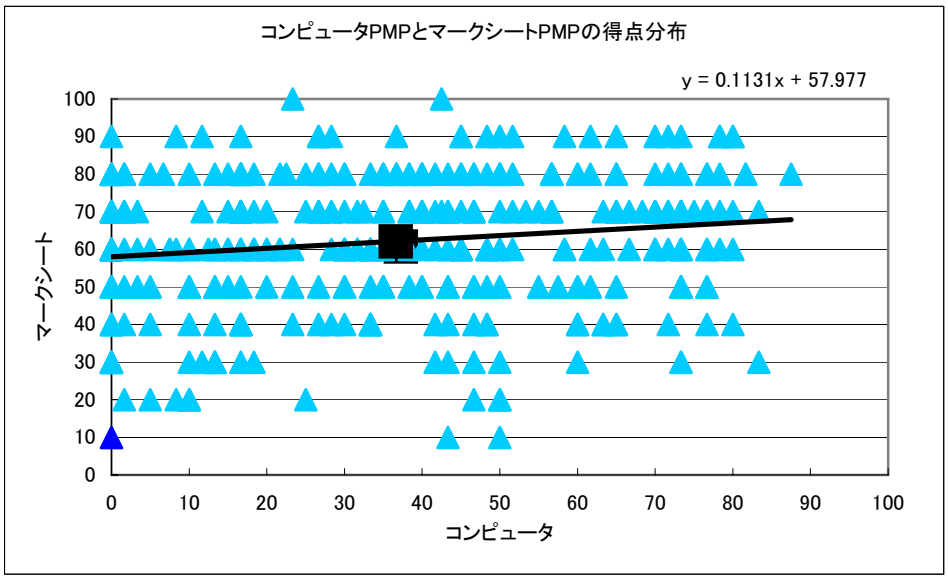
3題のＰＭＰ問題の得点とMCQの得点との間には強い相関はなかった(相関係数r=0.439)。男女比で見ると女性にやや相関がみられた。





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

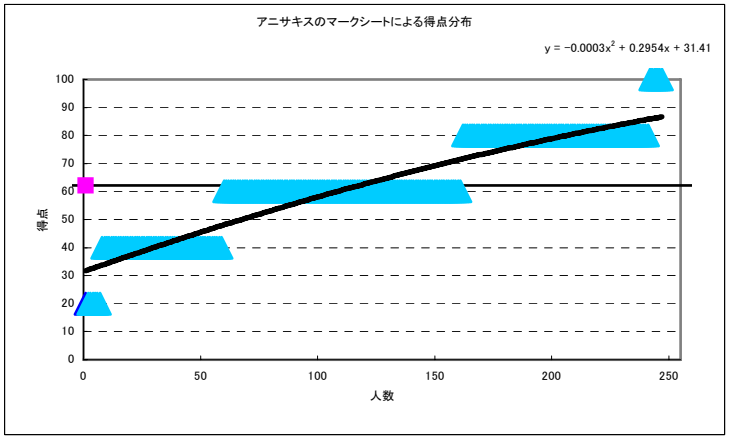
を問う試験問題になっていることから、PMP と異なった相関を示すものと考えられる。

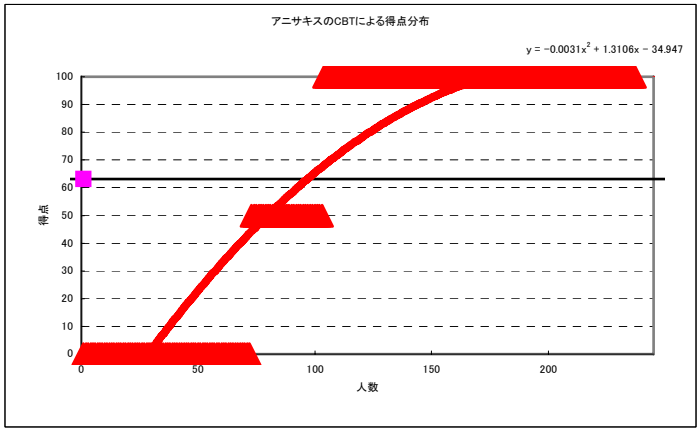


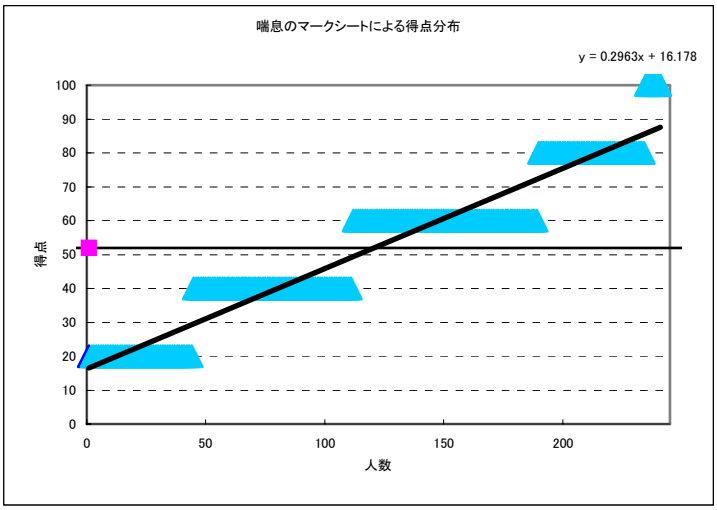
また PMP のコンピュータ問題とペーパー試験の直接の対比をみると、この散布図に示すように、両者間には相関がなかった。

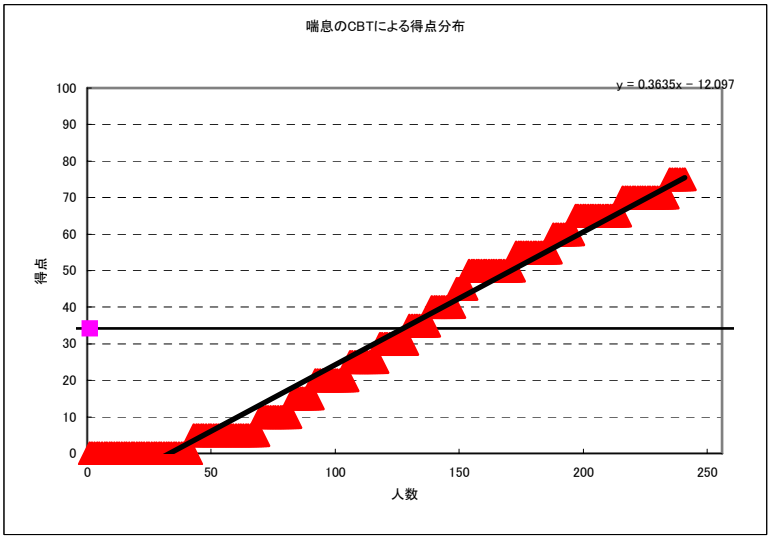
４．２　各 PMP 問題での検討

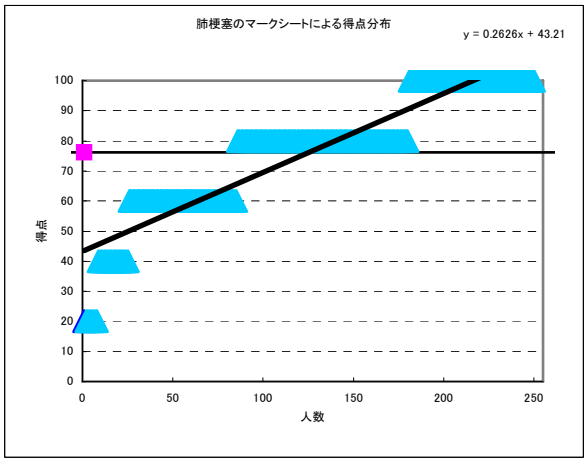
問題ごとにみていくとそれぞれ別の傾向がある。アニサキスでは回答項目が 3 項目で得点もその 3 つに分かれる。特に中間層は少なく、全然できないか、全て答えたッ科どちらかがほとんどになる。一方喘息は回答項目(治療法)が多種であり、得点分布もこれによって多数に分かれている。肺梗塞も治療法が多くあるために得点分布も広がるが、途中に大きなギャップがある。イレウスもできるものとできないものの差が明確になっており、かなりできるグループを入れると三群に分かれる。

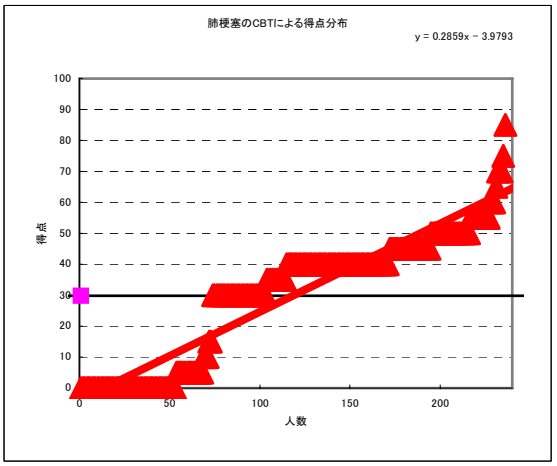

アニサキスのマークシートによる得点分布
y = −0.0003x² + 0.2954x + 31.41
得点
人数

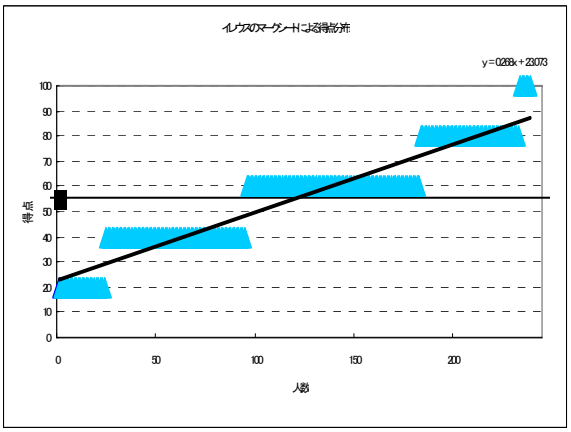

アニサキスのCBTによる得点分布
y = −0.0031x² + 1.3106x − 34.947
得点
人数

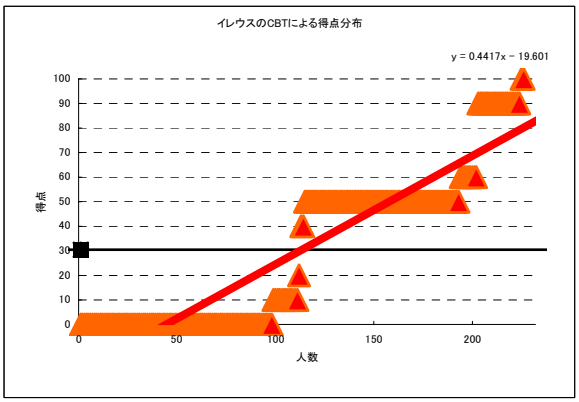

喘息のマークシートによる得点分布
y = 0.2963x + 16.178
得点
人数


喘息のCBTによる得点分布
y = 0.3635x − 12.097
得点
人数


肺梗塞のマークシートによる得点分布
y = 0.2626x + 43.21
得点
人数
0
10
20
30
40
50
60
70
80
90
100
0
50
100
150
200
250


肺梗塞のCBTによる得点分布
y = 0.2859x − 3.9793
得点
人数
0
10
20
30
40
50
60
70
80
90
100
0
50
100
150
200


イレウスのマークシートによる得点分布
y=0.268x+23.073
得点
人数
0
10
20
30
40
50
60
70
80
90
100
0
50
100
150
200

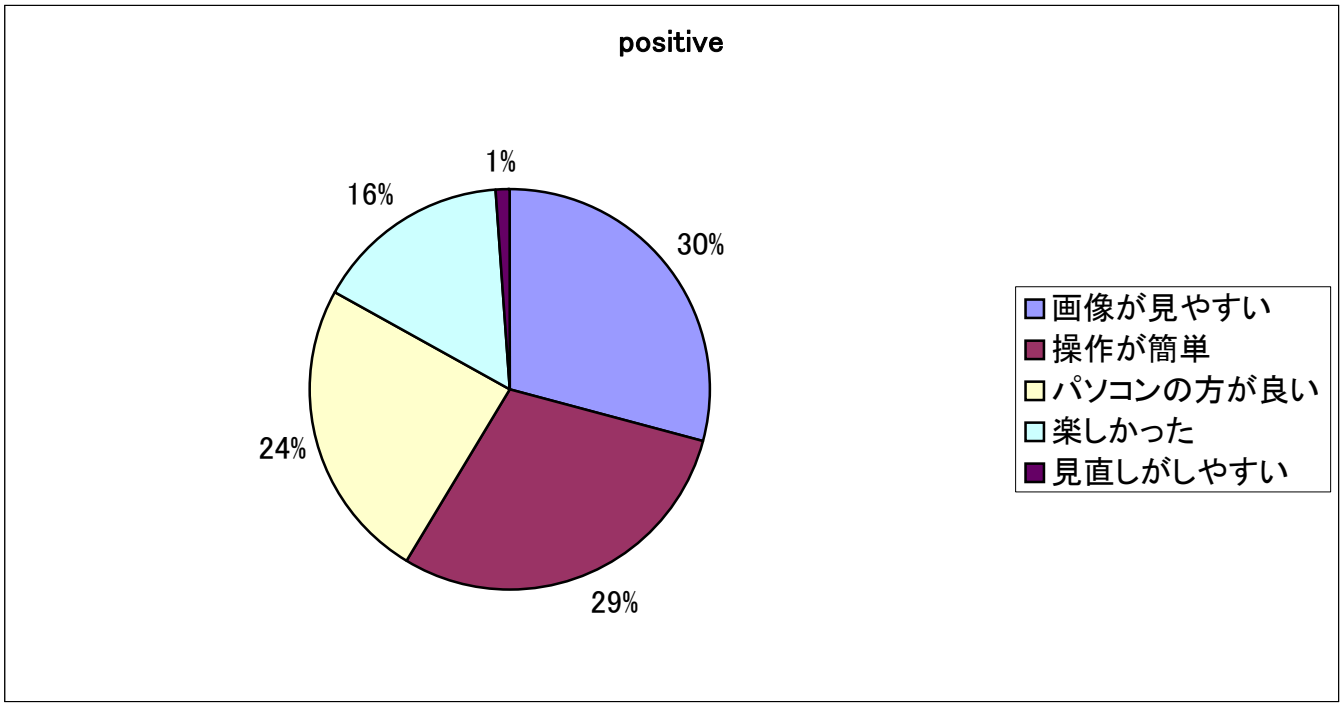

イレウスのCBTによる得点分布
y = 0.4417x − 19.601
得点
人数
0
10
20
30
40
50
60
70
80
90
100
0
50
100
150
200

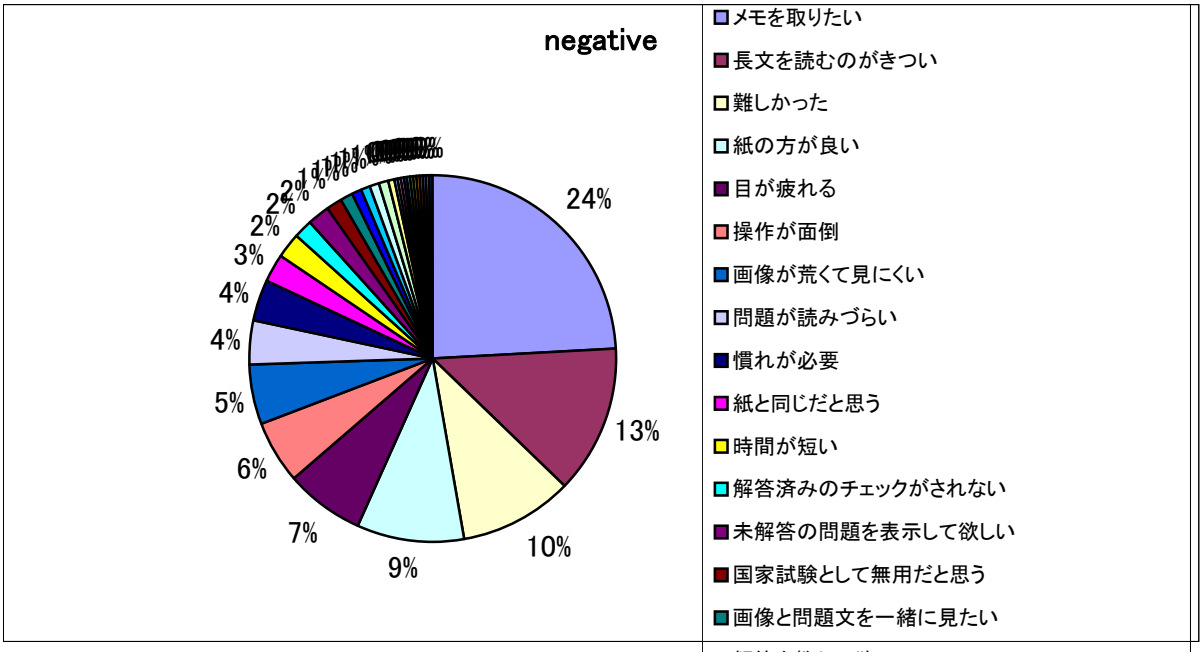
アンケート結果

1　MCQ

positive		negative	
画像が見やすい	24	メモを取りたい	92
操作が簡単	24	長文を読むのがきつい	50
パソコンの方が良い	20	難しかった	38
楽しかった	13	紙の方が良い	36
見直しがしやすい	1	目が疲れる	26
		操作が面倒	22
		画像が荒くて見にくい	20
		問題が読みづらい	14
		慣れが必要	14
		紙と同じだと思う	10
		時間が短い	8
		解答済みのチェックがされない	7
		未解答の問題を表示して欲しい	7
		国家試験として無用だと思う	5
		画像と問題文を一緒に見たい	4
		解答を教えて欲しい	4
		パソコンがエラーをおこす	3
		前の問題に戻りたい	3
		問題番号の表示をして欲しい	3
		文字が小さすぎる	2
		問題を返却して欲しい	1
		文字が大きすぎる	1
		動画がなく目新しさに欠けた	1
		図や写真は別冊にして欲しい	1
		終了時間になると勝手に閉じるようにして欲しい	1
		カウントダウン機能が欲しい	1
		絶望的	1
		システムがいまいち	1
		選択した画像をクリックしたら見れるようにして欲しい	1
		やりかたがわからない	1
		全画面表示にして欲しい	1
		希望者のみの受験にしてほしかった	1
		国試に準じた問題の方がよかった	1

表に 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。



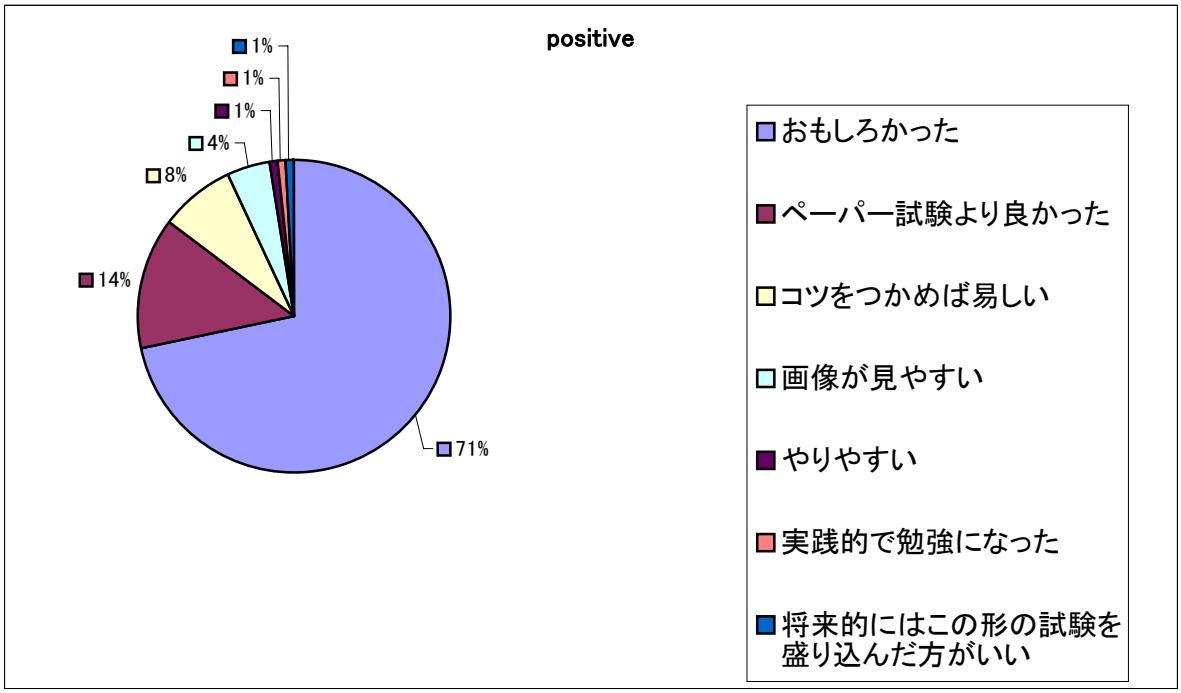


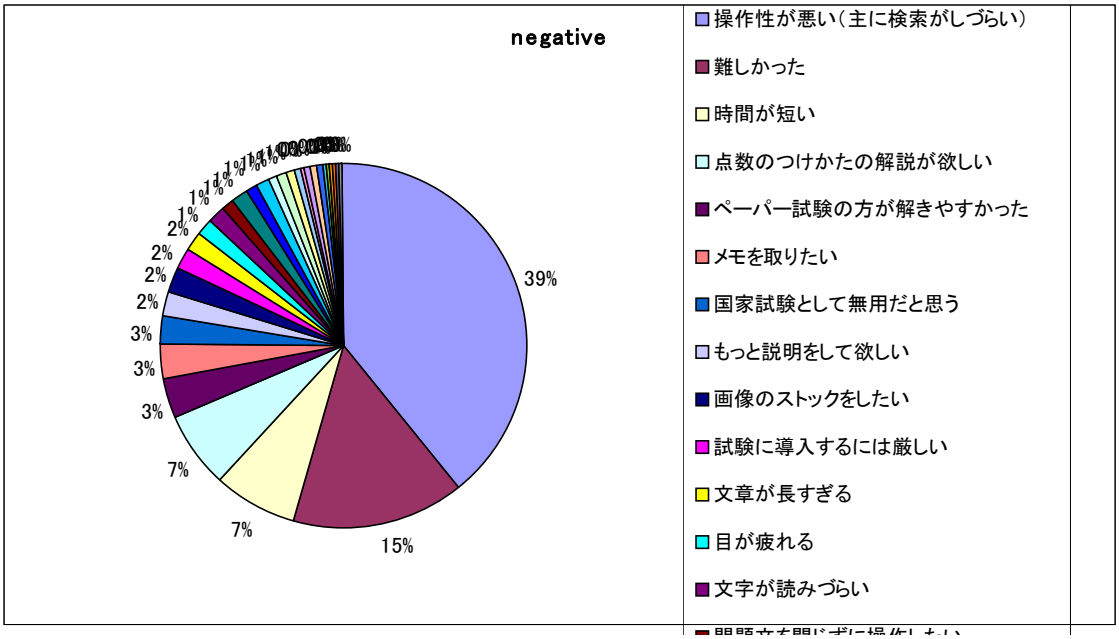
２．PMP

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。

positive		negative	
おもしろかった	83	操作性が悪い（主に検索がしづらい）	157
ペーパー試験より良かった	16	難しかった	61
コツをつかめば易しい	9	時間が短い	30
画像が見やすい	5	点数のつけかたの解説が欲しい	27
やりやすい	1	ペーパー試験の方が解きやすかった	14
実践的で勉強になった	1	メモを取りたい	12
将来的にはこの形の試験を盛り込んだ方がいい	1	国家試験として無用だと思う	11
		もっと説明をして欲しい	9
		画像のストックをしたい	8
		試験に導入するには厳しい	8
		文章が長すぎる	7
		目が疲れる	6
		文字が読みづらい	6
		問題文を閉じずに操作したい	5
		検査の項目を統一した方がよい	5
		普段の勉強から利用した方が良い	5
		カルテから写真が見たい	4
		パソコンがエラーをおこす	4
		画像が見にくい	3
		いつでも再アクセス出来るようにしてほしい	3
		文字を大きくして欲しい	2
		見にくい	2
		紙と一緒だと思う	2
		緊張した	2
		ゲーム感覚になり良くない	2
		もっと問題を多くして欲しい	1
		記載できる検査値の正常値がほしい	1
		パソコンが苦手な人には無理	1
		一長一短	1
		よくある症例を出して欲しい	1
		医学用語辞書を入れてほしい	1
		XP などで出せたら良かった	1


positive
1%
1%
1%
4%
8%
14%
71%
おもしろかった
ペーパー試験より良かった
コツをつかめば易しい
画像が見やすい
やりやすい
実践的で勉強になった
将来的にはこの形の試験を盛り込んだ方がいい


negative
39%
15%
7%
7%
3%
3%
3%
2%
2%
2%
2%
1%
1%
操作性が悪い（主に検索がしづらい）
難しかった
時間が短い
点数のつけかたの解説が欲しい
ペーパー試験の方が解きやすかった
メモを取りたい
国家試験として無用だと思う
もっと説明をして欲しい
画像のストックをしたい
試験に導入するには厳しい
文章が長すぎる
目が疲れる
文字が読みづらい

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、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。

考察

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差異があるわけではなかった。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、メモがとりたい、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。

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

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。

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。

今後

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。