資料　２－４

学力調査を活用した専門的な課題分析に関する調査研究

地域におけるデータ等を補完的に用いた調査分析手法の調査研究

平成21年度文部科学省企画公募委託研究

報告書

研究代表者　　石井　秀宗
（名古屋大学大学院教育発達科学研究科）

はじめに

本報告書は，文部科学省の企画公募委託研究「学力調査を活用した専門的な課題分析に関する調査研究」の『B．地域におけるデータ等を補完的に用いた調査分析手法の調査研究』に応募し，審査会を経て採用され，研究代表者らによって行われた研究の報告書である．

わが国においては，大規模テストは実施後，設問は公表されるがデータは公開されないため，テストを対象とした研究分野が発展しづらい状況にある．全国学力・学習状況調査についても，結果報告はあるものの（文部科学省・国立教育政策研究所，2007a，2007b，2008a，2008b，2009a，2009b など），他の研究者がそのデータを使って研究することはいまのところ難しい．これに対し欧米では，大規模テストのデータは原則として公開され（方法論上，設問は非公開とされることが多い），それを利用して誰でも研究を行うことができる環境にあり，テストに関する，または，テストデータを用いた研究が盛んに行われ，テスト研究分野の発展のみならず，教育実践，教育施策等にその知見が活かされている．

木村（2006）によれば，わが国において「テストの専門家」と言ったとき，戦後直後は「教科の専門家」「教育測定（テスト理論）の専門家」「サンプリングの専門家」を指していたが，やがて後二者が除かれ，代わりに「教育心理学者（教育評価の専門家）」が入り，その後，1979 年に共通第 1 次学力試験が導入される頃になると，大量のデータ処理の必要性から「情報処理の専門家」が加わり，近年は「教育社会学者」も「テストの専門家」を担うようになってきている．しかし，教育測定の専門家等が担うべき調査方法の設計や設問設計，テストの分析等については，依然として不十分なままでテストが実施されているのが，わが国の現状であると述べている．

こうした中，今回，初等中等教育を担う学校教諭と，教育測定学，計量心理学の研究者を含む研究組織により，全国学力・学習状況調査と地域におけるデータ等を補完的に用いた調査手法について研究する機会が設けられたことは貴重なことである．教育測定学的アプローチを用いればどのようなことが可能になるか，その一端を，短い期間ではあったが集中的に研究に取り組んだ成果として，ここに報告する．

2010 年 3 月

研究代表者　　石井　秀宗

研究組織

代表・統括	石井　秀宗
指導・助言	飯野　眞幸
	岡野　健
	野口　裕之
	柴山　直
	坂本　雄士
データ解析	安永　和央
	大羽　崇史
	鈴木　豪

研究期間

平成 22 年 1 月 15 日 ～ 平成 22 年 3 月 31 日

目　次

はじめに ……

１．序論 ……
(1)　学力の経年比較
(2)　本研究の目的
２．経年比較を行う方法 ……
(1)　記号の定義
(2)　データ収集のデザイン
(3)　得点の対応づけ
(4)　得点分布の推定と経年比較
３．得点の対応づけの方法 ……
(1)　テストの得点刻み
(2)　相対度数
(3)　分布関数
(4)　平滑化
(5)　パーセンタイル順位関数
(6)　パーセンタイル関数
(7)　等パーセンタイル法による対応づけ
４．対応づけを行う条件の検討 ……
(1)　テスト間相関
(2)　サンプルサイズ
(3)　テストの長さ
５．適用例 ……
(1)　データ
(2)　自治体テスト
(3)　全国学力・学習状況調査
(4)　テスト得点の分布
(5)　相対度数，分布関数，パーセンタイル順位関数
(6)　テスト得点の対応づけ
(7)　テスト得点分布の経年比較
６．データ収集デザインの拡張 ……
(1)　両地域において経年比較を行う場合
(2)　自治体テストが複数年度実施される場合
(3)　全国学力・学習状況調査が複数年度実施される場合
(4)　全国学力・学習状況調査が単年度に複数版実施される場合
(5)　地域が多数ある場合
(6)　拡張デザインと実際との対応
７．まとめ ……

文献 ……

1. 序論

(1) 学力の経年比較

学力低下とは，決して個人の学力が低下することを指しているのではない．少なくとも学齢期においては，個人の学力は向上こそすれ低下することは通常ない．学力低下という言葉は，年齢層が同等である過去の集団と現在の集団とを比較して，集団として学力が低下していることを指しており，経時的変化，通常は年単位の経年変化に着目して用いられている．

学力の経年比較は，原理的には，同一のテストを異なる時点で同等の集団に対して実施し，結果を比較することにより可能となるが，わが国においては，テスト問題は事後に公開され，また，公的な大規模テストになるほど試験対策指導が行われるなどという独特な試験文化があり（柳井・石井, 2008；柴山, 2008 など参照），一度実施したテストを別の時点で実施しても，適切に学力の経年変化を捉えられるとは言い難い状況にある．

例えば，同一の設問を異なる年度のテストに入れ，正答率を比較して経年比較を行うことにしても，少なくとも 2 つの問題点がある．それらは，教育測定学的には測定の妥当性の問題である．

1 つめの問題点は，最初に実施された後に問題が公開されていることから，次に受検する者は過去問題としてその問題を解いたことがあるか，または，試験対策としてその問題の学習が入念に行われ，その設問に対する準備だけはできた（他の問題だったらできなかった）可能性があるということである．このような状況で当該設問の正答率の比較をしたとしても，それは試験対策の効果または暗記力を検証しているだけに過ぎないという疑念が残る．

2 つめの問題点は，単一またはごく少数の設問だけでは，当該学習領域全般に関する議論はできないという問題である．ある特定の漢字の読み書きやことわざの意味がわかる（わからない）ことだけで「国語」の学力を論ずることは直感的には不合理なことと思われる．これは，入試問題や定期テストが，学習領域全体から満遍なく複数の設問で構成されていることを考えれば思い至ることである．受検者の学力を，様々な領域から，複数の項目を用いて，包括的に測定しているのである．学力の経年変化を捉える際も，多くの学習領域を網羅したテストを用いて学力を測定する必要がある．少数の設問で経年比較をしても，領域全体として適切な議論をしているかどうかは分からないのである．

これらの問題を回避する方法の 1 つとして，本実施の数年前に経年比較用のデータを収集しておくことが考えられる．比較検討に耐えるため，一定程度以上（具体的には数千人単位）の大きさのデータを収集する必要がある．しかし，公的なテストになるほど，本実施の前に（数千人規模の）調査を行っておくことは問題漏洩の観点から難しく（そのテストが経年比較に使われることが分かればなおさらである），この方法による経年比較は事実上不可能である．

また，別の解決策として，過去の集団でテストを実施していたら得られたであろう平均値や標準偏差などの分布パラメタの値を予想し，それらの値と現在の実際の値とを比較することも考え得るが，過去の値をどのように推定するのか，その適切性をどのように保証するかという新たな問題が生じてくる．なお，項目単位で目標正解率のようなものを設定した場合も，同じ問題が生じる．

よって，わが国においては，公的な大規模テストで学力の経年比較を行うことは非常に困難なことであると言える．これは，全国学力・学習状況調査についてもあてはまることである．

(2) 本研究の目的

しかし，教育測定学の領域で研究されている等化（Equating）や対応づけ（Linking）の理論を適用すれば，公的な大規模テストと，別に作成・実施されたテストとを適切に組み合わせて実施・分析することにより，項目単位の比較や予想値などとの比較よりもはるかに適切に，学力の経年変化を捉えることが可能となる．

そこで本研究では，全国学力・学習状況調査と，自治体等により作成・実施されたテストと組み合わせることにより，学力（具体的にはテストの得点分布）の経年比較を行う方法を提案し，その有効性を実際のデータを用いて提示することを目的とする．これは，全国学力・学習状況調査と自治体等が実施する学力調査を相補的に活用する 1 つの可能性を示すものであり，双方にとって経年比較が可能になるという利点を有するものである．

2. 経年比較を行う方法

(1) 記号の定義

議論を一般化するため，ここでは記号を用いて手法の説明を行う．

時点を表す記号を T_1,T_2，地域を表す記号を A,B，テストを表す記号を X,Y とする．テスト X とテスト Y は，同じ母集団を対象とし，同じ構成概念を測定すると見なせる同等なテストであるとする．また，各時点，各地域における測定は妥当なものであるとする．

(2) データ収集のデザイン

時点 T_1 は地域 A において，時点 T_2 は地域 B においてテスト X が実施されており，また，時点 T_2 には地域 A 及び地域 B においてテスト Y が実施されているものとする（図 1)．ただし，時点 T_2 の地域 B におけるテスト X とテスト Y の受検者は，同一の集団であるとする．また，時点 T_1 に地域 A においてテスト X を受検した集団と，時点 T_2 に地域 A においてテスト Y を受検した集団は同等な集団であるとする．地域 A における受検者集団と地域 B における受検者集団とが同等である必要は必ずしもないが，地域 A,B ともに妥当な測定がなされていなければならないので，同等な集団であるほうが望ましい．

以下，地域 A における時点 T_1 と時点 T_2 の，テスト X およびテスト Y の得点分布の経年比較を行う方法について述べていく．

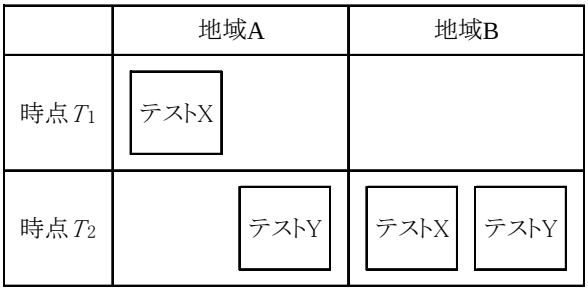


図1　データ収集デザイン

(3) 得点の対応づけ

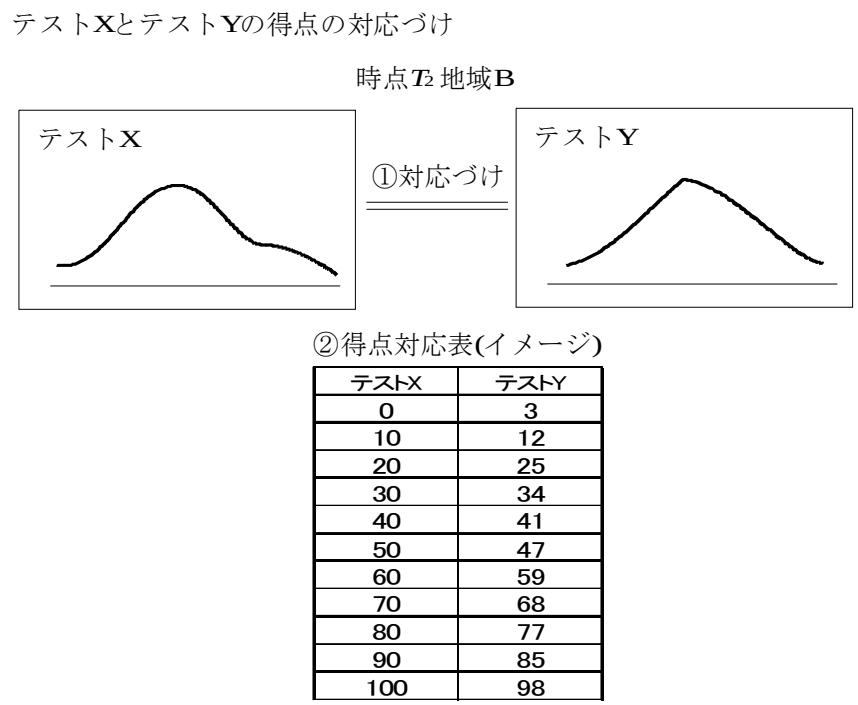
時点 T_2 に地域 B において実施されたテスト X とテスト Y は同一集団に対してなされたものであるから，2 つのテストの得点分布には対応関係があると考えられる（図 2 の①．以下同様)．この対応関係とは，それぞれの得点分布において相対的に同じところに位置する得点は等しい能力水準を表していると考えられる，というものである．この関係を「対応づけ（Linking)」の手法を用いて対応表にまとめ（②)，相互の得点を換算できるようにする．本研究で用いるテスト得点の対応づけの方法は次節で詳述する．

なお，もし地域 B の個々の受検者のテスト X とテスト Y の得点を知ることができれば（どの得点がどの受検者のものかをすべて特定することができれば)，得点間の相関係数を算出し両テストの 1 次元性を検証したり，等化の標準誤差をより正確に推定したりすることができる．しかし，逆に言えば，2 つのテストが同等な能力を測定していると見なすことに十分な根拠があれば，得点間の相関係数の値は大きいことが予測され，個人を特定する必要性は低くなる．よって，地域 B におけるテストの実施にあたっては，両方のテストを同一集団に対して実施しているという条件だけ満たせばよいことになる．これは，例えば学校単位，地域単位というまとまりが特定されていれば，個人までを特定する必要は必ずしも無いということを意味しており，データ収集デザイン上，非常に有効なことである．

(4) 得点分布の推定と経年比較

地域 A において，時点 T_1 にはテスト X が実施されているので，時点 T_1 のテスト X の得点分布は実在する．一方，時点 T_2 にはテスト Y は実施されているがテスト X は実施されていない．よって，時点 T_2 のテスト Y の得点分布は実在するがテスト X の得点分布は実在しない．

そこで，(3)で得られた対応表を用いてテスト Y の得点をテスト X の得点に換算して，時点 T_2 のテスト X の得点分布を推定する（③）．こうして得られた得点分布を用いれば，地域 A における時点 T_1 と T_2 のテスト X の得点分布の経年比較を行うことが可能となる（④）．



テストX	テストY
0	3
10	12
20	25
30	34
40	41
50	47
60	59
70	68
80	77
90	85
100	98

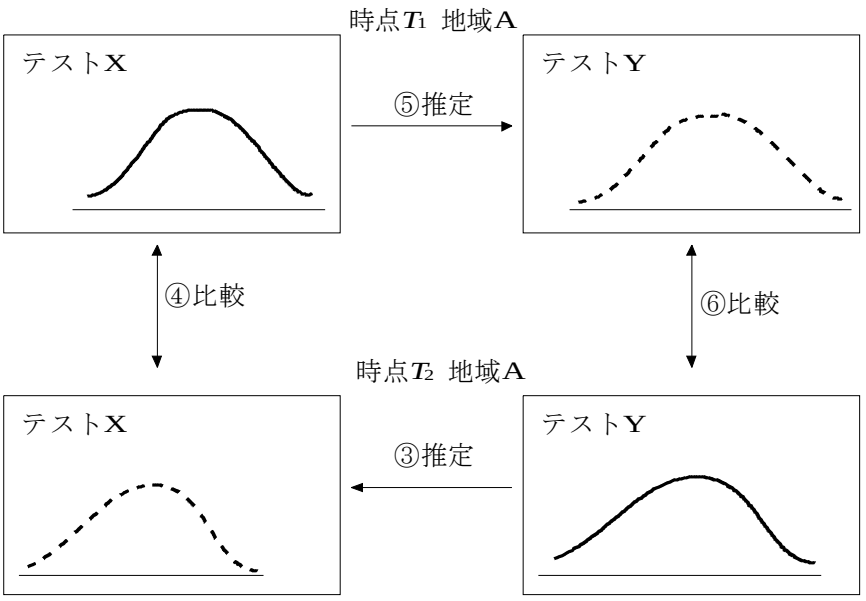


図2　得点の対応づけ，得点分布の推定，経年比較の様子

同様に，地域 A において，時点 T_2 にはテスト Y が実施されているので，時点 T_2 のテスト Y の得点分布は実在する．一方，時点 T_1 にはテスト X は実施されているがテスト Y は実施されていない．よって，時点 T_1 のテスト X の得点分布は実在するがテスト Y の得点分布は実在しない．

そこで，(3)で得られた対応表を用いてテスト X の得点をテスト Y の得点に換算して，時点 T_1 のテスト Y の得点分布を推定する（⑤）．こうして得られた得点分布を用いれば，地域 A における時点 T_1 と T_2 のテスト Y の得点分布の経年比較を行うことが可能となる（⑥）．

ここで，テスト X を地域において実施されたテスト，テスト Y を全国学力・学習状況調査とすれば，地域において実施されたテストと全国学力・学習状況調査の双方の得点分布について，地域 A における時点 T_1 と T_2 の経年比較を行うことが可能となる．すなわち，各自治体が作成・実施するテストと全国学力・学習状況調査とを相補的に用いて，学力の経年比較を行うことが可能となる．

3. 得点の対応づけの方法

テスト得点の対応づけ（または等化）にはいくつかの手法が考案されているが（Kolen & Brennan, 2004；前川, 1999 など参照），本研究では，比較的適用しやすく，実際のテストでも利用されている等パーセンタイル法（等百分位法）を用いることにする．

国内におけるテスト得点の経年変化を扱った先行研究としては，斉田(2003)や吉村他(2005)などがある．それらは多肢選択形式の設問からなる英語テストの得点の変化を，項目応答理論（Item Response Theory, IRT）を用いて分析している．項目応答理論を用いた場合は，各項目の特性値（パラメタ）やテスト情報量などを推定することができ，より多くの情報を導くことができるが，テストの得点分布の経年変化を捉えるだけであるならば，等パーセンタイル法などの古典的な手法でも十分対応でき，また解釈も容易である．

以下，柴山・野口(2004)を参考に，等パーセンタイル法によるテスト得点の対応づけの方法を説明する．基本的な考え方は，2 つのテストの得点分布において，相対的に同じところに位置する得点は等しい能力水準を表すと考えられるというものであり，概念的には図 3 のように理解される．

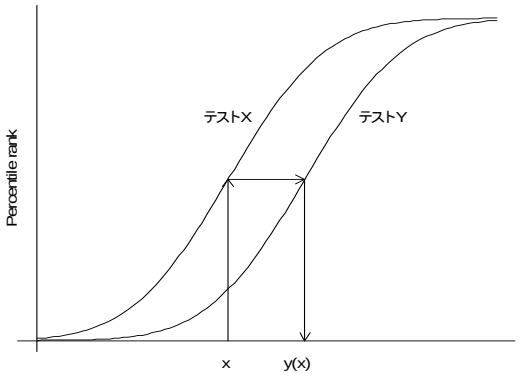


図3　等パーセンタイル法の概念図

(1) テストの得点刻み

テスト X において得点可能な値を 0,1,2,…,M_x とする．M_x はテスト X の満点である．得点刻みは均等に 1 点刻みとして説明するが，理論的には 1 点刻みでも均等（等間隔）である必要もない．しかし，実際のテストに適用する際には分布関数の平滑化（後述）をする場合があることなどを考えると，得点刻みは少なくとも等間隔であるのが望ましいと考えられる．

得点刻みが等間隔になるのは，各設問を正答・誤答の 2 値（1 と 0）で評価する場合，部分正答を認めるが評価点は等間隔である場合（例えば，正答=2，部分正答=1，誤答=0）などである．設問ごとに配点が異なったり部分得点の値が等間隔になっていない場合でも，多くの設問を組み合わせることにより得点刻みを等間隔にすることは原理的には可能であるが，このような場合，ある得点になる解答パターン数の違いにより，得点分布のところどころが煙突のように突出したり，針山状になったりする可能性が高くなる．そのような場合は，分布関数がぎくしゃくした形になるので，平滑化を行うことが望ましい．

(2) 相対度数

テスト X の全受検者数を N_x，得点が x 点であった受検者の人数を n_x とすると，得点が x 点の受検者の相対度数 $f(x)$ は，

$$f(x) = \frac{n_x}{N_x}, \quad x = 0, 1, 2, \ldots, M_X$$

と書ける．相対度数を 100 倍した値 100 $f(x)$ は，相対パーセントと呼ばれることがある．

(3) 分布関数

テスト X の得点が 0 点（最小点）から x 点までの相対累積度数 $F(x)$ は，

$$F(x) = \sum_{k=0}^{x} f(k)$$

と定義され，分布関数と呼ばれる．ただし，

$$\begin{cases} 0 \leq F(x) \leq 1, & x = 0, 1, 2, \ldots, M_X \\ F(x) = 0, & x < 0 \\ F(x) = 1, & x > M_X \end{cases}$$

である．相対累積度数を 100 倍した値 100 $F(x)$ は，累積パーセントと呼ばれることがある．

(4) 平滑化

部分点を付与する項目と付与しない項目があるなど，各項目の評価点の付け方が均一でない場合は，度数分布が針山状になり，分布関数がぎくしゃくした形になってしまうことがある．また，そもそも標本は誤差を含むものであるから，標本から得られた分布関数 $F(x)$ は誤差を含んでおり，それをそのまま用いて得点の対応づけを行うことは好ましくない場合もある．

そこで，得点の対応づけにあたって平滑化（スムージング）というプロセスが考案されている（Kolen & Brennan, 2004 など）．スムージングの手法は事前スムージングと事後スムージングに大別される．事前スムージングは分布関数を平滑化するものであり，事後スムージングは得点の対応関数を平滑化するものである．

平滑化の方法として簡便なのは移動平均法であるが，Kolen (1991) などの研究によると，移動平均法による平滑化では得点の対応がぎくしゃくしたり，得点段階数が少ない場合には不合理な対応を導く場合があることがなどが指摘されており，それに代わって，事前スムージングの方法としては，多項対数線形法（Polymomial Log-Linear Method）や，強真値法（Strong True Score Method），事後スムージングとしては 3 次スプラインによる対応関数の平滑化などの手法が開発されている（Kolen & Brennan, 2004）．

しかし，多項対数線形法などは多くのパラメタの値を推定しなければならないこと，経験的に簡便なのは移動平均法であること（前川， 1999），本研究では個々の受検者の得点ではなく受検者全体の得点分布の推定が行えればよいことなどから，ここでは移動平均法による平滑化の方法を説明する．

平滑化後の分布関数を $Fs(x)$ とすると，移動平均法による平滑化分布関数は，

$$F_s(x) = \sum_{k=x-h}^{x+h} w_k F(k), \quad x = h, h+1, \ldots, M_X - h, \quad \sum_{k=x-h}^{x+h} w_k = 1$$

により得られる．上式において，h=1 かつ w_k=1/3 とすれば，それは前後あわせて 3 つのデータの平均を取ることに相当し，次式のようになる．

$$\begin{cases} F_s(x) = \displaystyle\sum_{k=x-1}^{x+1} \frac{1}{3} F(k), \quad x = 1, 2, \ldots, M_X - 1 \\ F_s(0) = F(0) \\ F_s(M_X) = F(M_X) \end{cases}$$

(5) パーセンタイル順位関数

得点の低いほうから数えたときの各得点の相対的な順位をパーセントで表したものをパーセンタイル順位（percentile rank, PR）という（芝・南風原, 1990）. 例えば, 30 パーセンタイル順位というのは, その位置（得点）より下に全体の 30%のデータが存在することを表す. 同様に, 50 パーセンタイル順位というのは, その得点より下に全体の 50%, つまり半分のデータが存在することを表す.

テスト X の得点分布におけるパーセンタイル順位関数 $P(x)$ は次のように定義される. ただし, 式中にある x^*は, x に最も近い整数値を表す. つまり, この定義において x は整数である必要はなくなっている.

$$P(x) = \begin{cases} 100\Big(F(x^*-1) + \big(x - (x^* - 0.5)\big) \times \big(F(x^*) - F(x^*-1)\big)\Big), \\ \qquad -0.5 \le x \le M_X + 0.5 \\ 0, \quad x < -0.5 \\ 100, \quad x > M_X + 0.5 \end{cases}$$

もし, x が 1 点刻みであれば上式は,

$$P(x) = \begin{cases} 100\Big(F(x-1) + \dfrac{f(x)}{2}\Big), & x = 0, 1, 2, \ldots, M_X \\ 0, & x < 0 \\ 100, & x > M_X \end{cases}$$

と簡略化される.

ここで 0.5 や 1/2 などという数値が出てくるのは, 離散的な得点分布はもともとは連続的な値を区切ったものであると考えることによる. 例えば, 42 点という得点を得る受検者は, 41.5 点から 42.5 点の間に均等に散らばっていると考え, 本当に 42 点である受検者は, 42 点を観測した受検者の中でちょうど半分のところに位置すると考える. このような考えに従ってパーセンタイル順位を用いることの利点は, 得点の低いほうからの相対的な順位と, 得点の高いほうからの相対的な順位が対称性になることである（芝・南風原, 1990）.

なお, 分布関数 $F(x)$ を平滑化して $Fs(x)$ を作成した場合には, $F(x)$ の代わりに $Fs(x)$ やそれに対応する相対度数 $fs(x)$,

$$f_s(x) = F_s(x) - F_s(x-1)$$

を用いる.

パーセンタイル順位が分かっていないある得点 x について, x より小さい値の中でパーセンタイル順位が分かっている最大の x の値を x_L, x より大きい値の中でパーセンタイル順位が分かっている最小の x の値を x_U とする. このとき, x のパーセンタイル順位を線形補完法で推定するとすれば, その値は次式で算出される.

$$P(x) = \frac{P(x_U) - P(x_L)}{x_U - x_L}(x - x_L) + P(x_L) = \frac{P(x_U) - P(x_L)}{x_U - x_L}(x - x_U) + P(x_U)$$

この式は，テスト X の得点に対応づけられたテスト Y の得点がテスト X の得点可能な値でない場合に, 得点可能な値にそろえて得点の度数分布のグラフを書くときなどに役立つものである．

(6) パーセンタイル関数

あるパーセンタイル順位に対応するデータの値をパーセンタイルという（芝・南風原, 1990）．例えば，30 パーセンタイル順位に対応するデータの値が 42 点だったとしたら，この分布の 30 パーセンタイルは 42（点）ということになる．50 パーセンタイルは中央値（Q_2），25 パーセンタイルは第 1 四分位数（Q_1），75 パーセンタイルは第 3 四分位数（Q_3）と呼ばれる．

パーセンタイル順位関数 $P(x)$ の逆関数 $P^{-1}(x)$ をパーセンタイル関数と呼ぶ．パーセンタイル関数は，あるパーセンタイル順位 p が与えられたとき，それに対応する得点 x（パーセンタイル）を返すものであり，2 通りの定義式がある．

定義式1

$0 \leqq p < 100$ であるとき，累積パーセント $100F(x)$ の値が p 以上になる最小の整数値 x を x_U^* とすると，パーセンタイル関数 $P_U^{-1}(p)$ は,

$$x_U \equiv P_U^{-1}(p) \;=\; \frac{p/100 - F(x_U^* - 1)}{F(x_U^*) - F(x_U^* - 1)} + (x_U^* - 0.5)$$

と定義される．なお，p=100 の場合には,

$$x_U \equiv P_U^{-1}(100) \;=\; M_X + 0.5$$

と定義する．

定義式2

$0 < p \leqq 100$ であるとき，累積パーセント $100F(x)$ の値が p 以下になる最大の整数値 x を x_L^* とすると，パーセンタイル関数 $P_L^{-1}(p)$ は,

$$x_L \equiv P_L^{-1}(p) \;=\; \frac{p/100 - F(x_L^*)}{F(x_L^* + 1) - F(x_L^*)} + (x_L^* + 0.5)$$

と定義される．なお，p=0 の場合には,

$$x_L \equiv P_L^{-1}(0) \;=\; -0.5$$

と定義する．

もし，すべての x について $f(x) > 0$ なら $x_U = x_L\ (= x)$ となるため，どちらの定義式を用いても構わない．しかし，$f(x)$=0 なる x が存在する，すなわち，当該の得点となる受検者がいない値がある場合には，$x_U \neq x_L$ となる．このような場合は,

$$x \;=\; \frac{x_U + x_L}{2}$$

とするのが自然であると考えられる．

実際にパーセンタイルを求める際には，常に x_U と x_L の両方を求めその平均を取るようにしておくと, $f(x)$=0 となる x が存在するかどうかを気にせずにパーセンタイルを求めることができる．

(7) 等パーセンタイル法による対応づけ

テスト X とテスト Y の 2 つのテストがあるとき，テスト X においてあるパーセンタイル順位をもつ得点 x に対し，テスト Y においてそのパーセンタイル順位に対応する得点（パーセンタイル）$e_Y(x)$ を対応させる関数を，テスト X のテスト Y への対応関数と呼ぶことにする．

テスト X と同様に，テスト Y の得点可能な値を 0,1,2,…,M_Y，相対度数を $g(y)$，分布関数を $G(y)$，パーセンタイル順位関数を $Q(y)$，パーセンタイル関数を $Q^{-1}(y)$ とすると，$e_Y(x)$ は，

$$e_Y(x) = Q^{-1}(P(x)), \quad -0.5 \le x \le M_X + 0.5$$

で定義できる．もし $g(y)=0$ となる y が存在すれば，パーセンタイル関数を定義したときと同様に，2 通りの定義ができる．

対応関数1

$$e_{Y_U}(x) = Q_U^{-1}(P(x)) = \begin{cases} \dfrac{P(x)/100 - G(y_U^* - 1)}{G(y_U^*) - G(y_U^* - 1)} + (y_U^* - 0.5), & 0 \le P(x) < 100 \\ M_X + 0.5, & P(x) = 100 \end{cases}$$

対応関数2

$$e_{Y_L}(x) = Q_L^{-1}(P(x)) = \begin{cases} \dfrac{P(x)/100 - G(y_L^*)}{G(y_L^* + 1) - G(y_L^*)} + (y_L^* + 0.5), & 0 < P(x) \le 100 \\ -0.5, & P(x) = 0 \end{cases}$$

$g(y)=0$ となる y が存在しなければ $e_{YU}(x)$ と $e_{YL}(x)$ は一致するが，$g(y)=0$ となる y が存在すれば $e_{YU}(x)$ と $e_{YL}(x)$ は一致しない．そこで，常に $e_{YU}(x)$ と $e_{YL}(x)$ の両方を求めその平均を取るようにしておくと，$g(y)=0$ なる x が存在するかどうかを気にせずに $e_Y(x)$ を求めることができる．よって，テスト X の得点 x に対応するテスト Y の得点は，

$$e_Y(x) = \frac{e_{Y_U}(x) + e_{Y_L}(x)}{2}$$

と計算される．なお，分布関数 $G(x)$ を平滑化して $Gs(x)$ を作成した場合には，$G(x)$ の代わりに $Gs(x)$ を用いる．

テスト Y の得点 y に対応するテスト X の得点を算出する際は，上記の議論において，テスト X とテスト Y に関する項をすべて逆転させればよい．

４．対応づけを行う条件の検討

得点の対応づけは，技術的には任意のテスト間の対応づけが可能であるが，解釈可能な結果を得るためには，ある程度の条件を満たしている必要がある．本節では，等パーセンタイル法を用いてテスト得点の対応づけを行う際，テスト間にどの程度の相関があればよいか，得点対応表を作成するために必要なサンプルサイズはどれくらいか，テストの長さはどれくらいであるべきかなど，実際に問題となってくる事項について検討する．

(1) テスト間相関

異なるテスト仕様のもとで開発された 2 つのテスト得点について，得点を直接的に対応づけられるようにするためには，測定している概念，信頼性，及び，困難度が，それぞれ似通っている必要がある（Holland & Dorans, 2006, p193）．しかし，実際どの程度の相関があれば，この条件を満たしているかについての明確な基準はない．そこで，実際の適用例を見てみると，柴山・野口(2004)は，法科大学院統一適性試験と法科大学院適性試験が上記の状況にある 2 つのテストであるとし，相関係数を推定すると r=0.68 で，得点の対応づけが適用可能である判断している．また，Kolen & Brennan（2004, p.260)では，r=0.7 として対応づけの精度を評価する例が示されている．一方，同一のテスト仕様のもとで開発された平行テストの等化を行う際，平行テスト間の相関係数は r=0.9 が典型的な値であると言われている（Lord, 1982）．得点の対応づけは得点の等化よりも緩やかな条件のもとで行われることを考えれば，柴山・野口(2004)や Kolen & Brennan（2004)が用いた r=0.7 程度という値は，十分理解できる値である．

よって，得点の対応づけを行うにあたっては，テスト間相関としては，おそらく r=0.7 程度以上が 1 つの目安にはなると考えられ，似たような（または同一の）構成概念の測定を念頭におき，異なるテスト仕様のもとで開発されたテスト間の相関は実際その程度になることが多いと予想される．しかし，実際に対応づけを行う際，2 つのテストを受検した集団において個人が特定できなければ，実際のテスト間相関係数は計算できない．その場合は，測定している概念，測定の信頼性，困難度が似通っているかを検討するしかない．測定している構成概念については，テストの設計段階におけるテスト仕様を確認することになる．信頼性と困難度については，信頼性係数や平均項目正答率を推定することなどにより，数量的に検討することが可能である．

(2) サンプルサイズ

対応づけの精度は，得点の対応表がどの程度の精度で作成されているかによる．これは等化の標準誤差というもので評価することが可能である．2 つのテスト得点が相関係数 ρ の 2 変量正規分布に従うと仮定すると，同一集団に 2 つのテストを実施した際の等化得点 $e_Y(X)$ の標準誤差 se_z は近似的に次式のようになる（Lord, 1982 より）．

$$se_z^2 = \frac{2\gamma_z\sigma_Y^2}{Ng_z^2}$$

ただし，N はサンプルサイズ，σ_Y はテスト Y の得点分布の標準偏差，z はテスト X の得点を標準化した得点，g_z は得点 z における標準正規分布の密度関数の値，γ_z は相関係数 ρ の 2 変量標準正規分布において（$y < z$, $x > z$）となる確率を表す．

等化の標準誤差の大きさに対する基準として Kolen & Brennan（2004, p.258)は，標準誤差 se_z が標準偏差 σ_Y の 0.1 倍以下になることを用いている．これを満たすサンプルサイズ N_z を z の関数として導くと，その値は，u=0.1 として次式により推定される．

$$N_z = \frac{2\gamma_z}{u^2 g_z^2}$$

表 1 は，2 つのテスト間相関を ρ =0，0.5，0.7，0.8，0.9 とした場合，また，z=0，1，1.5，2，2.5 とした場合の N_z の値である．Kolen & Brennan（2004, p.288）は，z=2（平均の周りの 95.45%の範囲）における等化の標準誤差の割合を 0.1 以下にすることを，対応づけに必要なサンプルサイズを推定する基準として用いており，これに従うと，表 1 から，ρ =0.7 の場合は 1,056 名，ρ =0.9 の場合は 645 名の受検者が必要と推定される．Downing & Haladyna（2006，p544）においても対応づけに必要なサンプルサイズは千名以上と述べられている．

よって，同一集団に 2 つのテストを実施し等パーセンタイル法により得点の対応づけを行う際に必要なサンプルサイズはおよそ 1,000 名程度，テスト間相関がかなり高い場合（r=0.9）には 650 名程度，反対にテスト間相関がほとんど 0 に近い場合でも 1,500 名程度と推定される．

表1　対応づけに必要なサンプルサイズの推定

z	平均の周りの範囲	テストXとテストYの相関係数 ρ				
		0.0	0.5	0.7	0.8	0.9
0	0.00%	315	210	160	129	91
1	68.27%	456	329	256	209	148
1.5	86.64%	744	579	461	381	273
2	95.45%	1526	1283	1056	887	645
2.5	98.76%	4018	3607	3076	2631	1948

(3)　テストの長さ

テストの長さ（項目数）はどれくらいが適切かという問題について Brennan & Kolen（1987）は，得点の等化を行うテストの長さは 30 ～ 40 項目以上という基準を示している．この基準は，各項目が正答か誤答かの 2 値で採点される状況を念頭に置いていることから，部分点を与えるようなわが国のテストにおいては，得点段階が 30 ～ 40 段階以上と読み替えるのが適当であると考えられる．

しかし，テスト項目数は，テストの目的，測定する構成概念，実施時間などによっても異なるものであり，一概なことは言えない，とも述べられている（Kolen & Brennan，2004, p.270）．実際，Kolen & Brennan（2004, p.220）では，24 項目からなるテストの得点を等化している例も示されている．

得点の対応づけは，項目数が 2 や 3 の場合でも技術的には可能である．しかし，項目数が 2 や 3 であるテストで測定の精度は十分か，得点に具体的な解釈を与えることが適切か，具体的な解釈を与えることに意味があるかなどを考えると，少数項目からなるテスト得点の対応づけは，多くの課題を抱えていると言える．パフォーマンス評価得点の対応づけなどが，この状況にあてはまる（Kolen & Brennan，2004, p.322）．

技術的に可能である以上，得点の対応づけに必要なテストの長さについては，少なくとも通常のテスト作成で要求される項目数は必要であると言える．逆に言えば，項目数（得点段階数）が少なく，測定の精度や解釈可能性の問題があっても，他でその得点についての解釈を行うようなことをしているのであれば，対応づけに必要なサンプルサイズが確保されている限り得点の対応づけも認められる（認めざるを得ない）ということになろう．

5. 適用例

本節では，上述した方法を実際のデータに適用し，テストの得点分布の経年比較を行う．本研究の目的が，「全国学力・学習状況調査と，自治体等により作成・実施されたテストと組み合わせることにより，学力（具体的にはテストの得点分布）の経年比較を行う方法を提案し，その有効性を実際のデータを用いて提示すること」であることから，適用例としては，自治体等が作成・実施したテストと全国学力・学習状況調査を用いた例を示すことにする．

(1) データ

データの概略を図 4 に示す．自治体等が作成・実施したテスト（テスト X）として，平成 18 年 4 月（時点 T_1）に，ある自治体（地域 A）において実施され，また，平成 21 年 4 月頃（時点 T_2）に，5 つの学校（地域 B，以下，5 中学校）において実施された中学校 3 年生の国語のテストを用いることにする．地域 A における平成 18 年 4 月の有効解答者数は 3,892 名であり，平成 21 年 4 月頃の 5 中学校における有効解答者数は 778 名である．

テスト Y に相当するテストは，平成 21 年 4 月（時点 T_2）に実施された全国学力・学習状況調査の中学校国語テスト（全国学力・学習状況調査）である．地域 A での解答者数は 19,498 名であり，5 中学校での解答者数は 759 名である．

この適用例における得点の対応づけに用いる受検者数（759 名）は，仮にテスト X と全国学力・学習状況調査の相関係数が 0.7 であるとすると，平均の周りの約 93%の範囲で，標準偏差に対する等化の標準誤差の割合が 0.1 以下になる値であり，対応づけの精度は十分確保されていると考えられる．

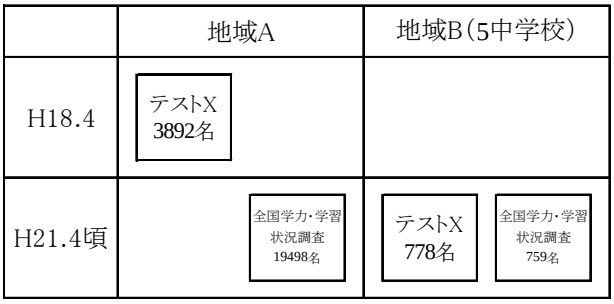


図4　データの概略

(2) 自治体テスト

自治体テスト（テスト X）は，説明文とそれに関する設問，記述式 9 問，穴埋め 3 問，選択式 6 問の合計 18 問からなる読解テストで，50 分で実施されるテストである．このテストは，大規模なデータが収集されており，また，受検者の動機づけも高いと考えられることなどの特質を持っており，全国学力・学習状況調査と対応づけを行うに十分適したテストであると考えられる．

本研究においては，5 中学校で同等に実施された 11 設問（設問 3,4,5,6,7,8,9,12,14,15,16）だけからなるテストを考えることにする．これら以外の設問は，平成 18 年度に地域 A で実施されたときと採点方法が異なるか，5 中学校で異なる形式を用いて実施され等価性が保たれていないかするため，今回の分析からは省く．5 中学校で同等に実施された 11 設問からなるテストを改めてテスト X とする．この設問の中には部分得点を付与する項目も含まれるので（0，0.25，0.5，1 など），テスト X の得点刻みは 0.25 点で，可能得点範囲は，0，0.25，0.50，…，10.75，11 点の 45 段階である．

(3) 全国学力・学習状況調査

平成 21 年度全国学力・学習状況調査の中学校国語のテストは，問題 A「知識」33 問，問題 B「活用」11 問の合計 44 問からなる．各設問は正答（1 点）か誤答（0 点）で採点される．

本研究では，全 44 問からなるテストをテスト M と表記する．テスト M の得点可能範囲は 0 ～ 44 点（45 段階）である．

問題 A をテスト MA，問題 B をテスト MB とする．それぞれのテストの得点可能範囲は，MA：0 ～ 33 点（34 段階），MB：0 ～ 11 点（12 段階）である．

また，課題領域として，「話すこと・聞くこと」に関する 4 設問，「書くこと」に関する 8 設問，「読むこと」に関する 21 設問，「言語事項」に関する 17 設問からなるテストをそれぞれ，テスト M1，M2，M3，M4 とする．それぞれのテストの得点可能範囲は，M1：0 ～ 4 点（5 段階），M2：0 ～ 8 点（9 段階），M3：0 ～ 21 点（22 段階），M4：0 ～ 17 点（18 段階）である．

以上，全国学力・学習状況調査からは，テスト M，MA，MB，M1，M2，M3，M4 の 7 つのテストが構成される．簡単のために，必要に応じてこれらをテスト M's と略記する．

(4) テスト得点の分布

5 中学校におけるテスト X とテスト M's の α 係数及び得点分布の基本統計量を表 2, テスト M's 間の相関係数を表 3 に示す．また分布の様子を図 5 に描く．α 係数は信頼性係数（池田，1994 など）の推定量の 1 つで，当該のテストに含まれる項目得点の内的整合性を反映するものである．値が 1 に近いほど測定の信頼性が高いと評価されるが，一般に項目数が大きくなると α 係数の値は大きくなる傾向にあること，少数項目で α 係数の値が高い場合は測定の妥当性が低い可能性があること（村上， 2008 など）に注意する必要がある．本研究においては，テスト M!が項目数が少なく，また α 係数が小さいので，テスト M1 に関する考察は，行うとすれば，慎重に扱うべきである．

表2　5中学校における各テスト得点分布の基本統計量

	項目数	α係数	人数	平均	SD	最小値	中央値	最大値
X	11	0.71	778	6.88	2.41	0.00	7.00	11.00
M	44	0.91	759	33.79	8.06	0	36	44
MA	33	0.88	759	25.47	5.92	0	27	33
MB	11	0.77	759	8.31	2.53	0	9	11
M1	4	0.45	759	3.43	0.84	0	4	4
M2	8	0.69	759	5.63	2.01	0	6	8
M3	21	0.86	759	15.82	4.34	0	17	21
M4	17	0.79	759	13.12	3.24	0	14	17

表3　5中学校におけるテストM's間の相関係数

	M	MA	MB	M1	M2	M3	M4
M	1						
MA	0.98	1					
MB	0.89	0.79	1				
M1	0.64	0.65	0.53	1			
M2	0.86	0.80	0.87	0.51	1		
M3	0.96	0.90	0.94	0.56	0.86	1	
M4	0.92	0.94	0.73	0.51	0.72	0.79	1

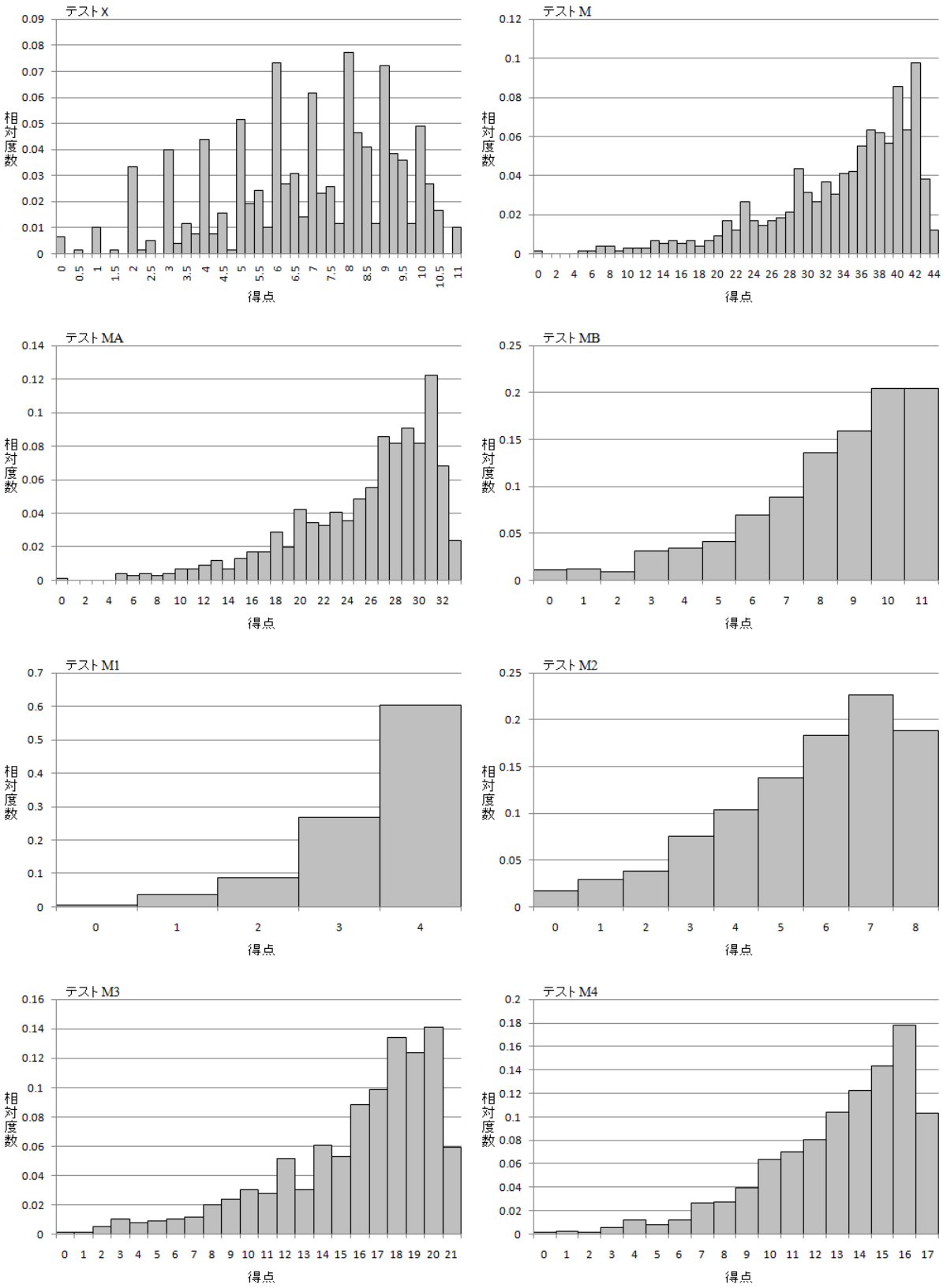


図5　5中学校におけるテストXとテストM'sの得点分布

(5) 相対度数，分布関数，パーセンタイル順位関数

平成 21 年 4 月頃に 5 中学校で実施されたテスト X と全国学力・学習状況調査の得点分布について，テスト X とテスト M's の相対度数 $f(x)$，分布関数 $F(x)$，パーセンタイル順位関数 $P(x)$ の値を算出する．テスト X は分布関数 $F(x)$ がぎくしゃくしているため，平滑化した分布関数 $Fs(x)$ を作成し，それを用いてパーセンタイル順位関数を求める．分布関数の平滑化にあたっては，前後をあわせた 3 つの値の平均を取る方法を用いる．この平滑化による系統的な歪みは生じていない．以上の結果を表 4 ～表 11，パーセンタイル順位関数のグラフを図 6 に示す．

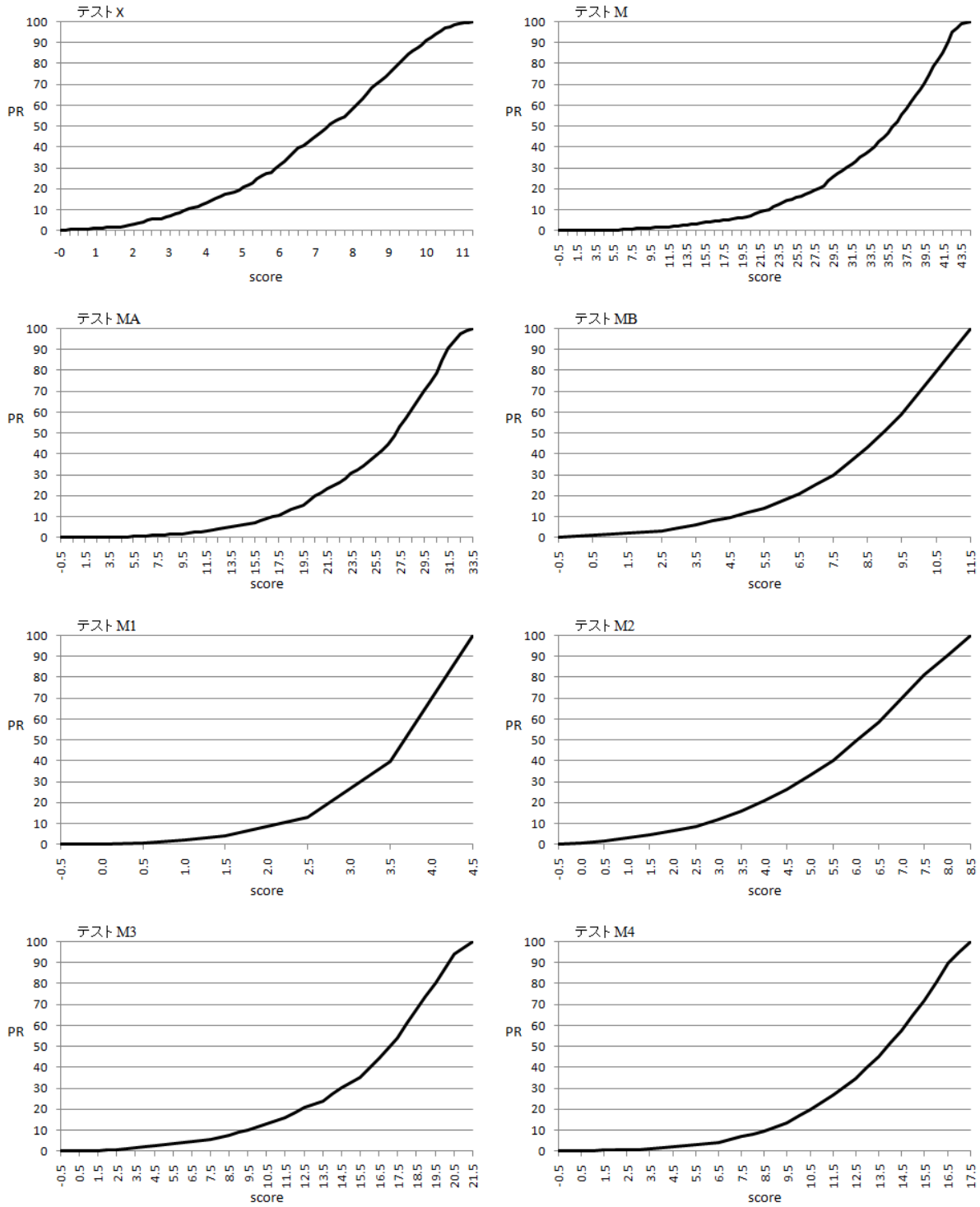


図6　5中学校におけるテストXとテストM'sのパーセンタイル順位関数のグラフ

表4　テストXの$f(x)$, $F(x)$, $Fs(x)$, $P(x)$

x	$f(x)$	$F(x)$	$Fs(x)$	$P(x)$
0.00	.0064	.0064	.0064	.32
0.25	.0000	.0064	.0069	.66
0.50	.0013	.0077	.0073	.71
0.75	.0000	.0077	.0111	.92
1.00	.0103	.0180	.0146	1.29
1.25	.0000	.0180	.0184	1.65
1.50	.0013	.0193	.0189	1.86
1.75	.0000	.0193	.0304	2.46
2.00	.0334	.0527	.0420	3.62
2.25	.0013	.0540	.0553	4.86
2.50	.0051	.0591	.0574	5.63
2.75	.0000	.0591	.0724	6.49
3.00	.0398	.0990	.0870	7.97
3.25	.0039	.1028	.1054	9.62
3.50	.0116	.1144	.1131	10.93
3.75	.0077	.1221	.1341	12.36
4.00	.0437	.1658	.1538	14.40
4.25	.0077	.1735	.1761	16.50
4.50	.0154	.1889	.1842	18.02
4.75	.0013	.1902	.2069	19.56
5.00	.0514	.2416	.2309	21.89
5.25	.0193	.2609	.2626	24.68
5.50	.0244	.2853	.2806	27.16
5.75	.0103	.2956	.3166	29.86
6.00	.0733	.3689	.3535	33.50
6.25	.0270	.3959	.3972	37.53
6.50	.0308	.4267	.4212	40.92
6.75	.0141	.4409	.4567	43.89
7.00	.0617	.5026	.4897	47.32
7.25	.0231	.5257	.5266	50.81
7.50	.0257	.5514	.5467	53.66
7.75	.0116	.5630	.5848	56.58
8.00	.0771	.6401	.6298	60.73
8.25	.0463	.6864	.6847	65.72
8.50	.0411	.7275	.7177	70.12
8.75	.0116	.7391	.7592	73.84
9.00	.0720	.8111	.7999	77.96
9.25	.0386	.8496	.8488	82.43
9.50	.0360	.8856	.8775	86.31
9.75	.0116	.8972	.9096	89.35
10.00	.0488	.9460	.9387	92.42
10.25	.0270	.9730	.9696	95.42
10.50	.0167	.9897	.9841	97.69
10.75	.0000	.9897	.9931	98.86
11.00	.0103	1.0000	1.0000	99.66

表5　テストMの$f(x)$, $F(x)$, $P(x)$

x	$f(x)$	$F(x)$	$P(x)$
0	.0013	.0013	.07
1	.0000	.0013	.13
2	.0000	.0013	.13
3	.0000	.0013	.13
4	.0000	.0013	.15
5	.0013	.0026	.22
6	.0013	.0040	.37
7	.0040	.0079	.64
8	.0040	.0119	.94
9	.0013	.0132	1.23
10	.0026	.0158	1.47
11	.0026	.0184	1.71
12	.0026	.0211	2.04
13	.0066	.0277	2.48
14	.0053	.0329	3.03
15	.0066	.0395	3.62
16	.0053	.0448	4.22
17	.0066	.0514	4.79
18	.0040	.0553	5.34
19	.0066	.0619	5.95
20	.0092	.0711	6.83
21	.0171	.0883	8.01
22	.0119	.1001	9.57
23	.0264	.1265	11.42
24	.0171	.1436	13.31
25	.0145	.1581	15.09
26	.0171	.1752	16.73
27	.0184	.1937	18.51
28	.0211	.2148	20.84
29	.0435	.2582	23.83
30	.0316	.2899	27.12
31	.0264	.3162	30.39
32	.0369	.3531	33.53
33	.0303	.3834	36.89
34	.0408	.4242	40.58
35	.0422	.4664	44.77
36	.0553	.5217	49.76
37	.0632	.5850	55.45
38	.0619	.6469	61.48
39	.0567	.7036	67.92
40	.0856	.7892	74.75
41	.0632	.8524	82.28
42	.0975	.9499	89.70
43	.0382	.9881	95.48
44	.0119	1.0000	98.97

表6　テストMAの$f(x)$, $F(x)$, $P(x)$

x	$f(x)$	$F(x)$	$P(x)$
0	.0013	.0013	.07
1	.0000	.0013	.13
2	.0000	.0013	.13
3	.0000	.0013	.13
4	.0000	.0013	.20
5	.0040	.0053	.37
6	.0026	.0079	.66
7	.0040	.0119	.99
8	.0026	.0145	1.32
9	.0040	.0184	1.71
10	.0066	.0250	2.22
11	.0066	.0316	2.88
12	.0092	.0408	3.71
13	.0119	.0527	4.63
14	.0066	.0593	5.62
15	.0132	.0725	6.76
16	.0171	.0896	8.17
17	.0171	.1067	10.01
18	.0290	.1357	12.17
19	.0198	.1555	14.78
20	.0422	.1976	17.90
21	.0343	.2319	21.32
22	.0329	.2648	24.95
23	.0408	.3057	28.57
24	.0356	.3412	32.48
25	.0487	.3900	36.89
26	.0553	.4453	42.38
27	.0856	.5310	49.25
28	.0817	.6126	57.27
29	.0909	.7036	65.81
30	.0817	.7852	74.97
31	.1225	.9078	84.43
32	.0685	.9763	92.56
33	.0237	1.0000	98.07

表7　テストMBの$f(x)$, $F(x)$, $P(x)$

x	$f(x)$	$F(x)$	$P(x)$
0	.0105	.0105	.53
1	.0119	.0224	1.60
2	.0092	.0316	3.03
3	.0316	.0632	5.16
4	.0343	.0975	8.19
5	.0408	.1383	12.38
6	.0698	.2082	18.12
7	.0883	.2964	26.33
8	.1357	.4321	37.62
9	.1594	.5916	52.33
10	.2042	.7958	70.11
11	.2042	1.0000	89.79

表8　テストM1の$f(x)$, $F(x)$, $P(x)$

x	$f(x)$	$F(x)$	$P(x)$
0	.0053	.0053	.26
1	.0356	.0408	2.31
2	.0883	.1291	8.50
3	.2675	.3966	26.28
4	.6034	1.0000	69.83

表9　テストM2の$f(x)$, $F(x)$, $P(x)$

x	$f(x)$	$F(x)$	$P(x)$
0	.0171	.0171	.86
1	.0290	.0461	3.32
2	.0382	.0843	7.29
3	.0751	.1594	13.29
4	.1041	.2635	22.20
5	.1383	.4018	34.58
6	.1831	.5850	50.81
7	.2266	.8116	69.92
8	.1884	1.0000	89.94

表10　テストM3の$f(x)$, $F(x)$, $P(x)$

x	$f(x)$	$F(x)$	$P(x)$
0	.0013	.0013	.07
1	.0013	.0026	.26
2	.0053	.0079	.68
3	.0105	.0184	1.36
4	.0079	.0264	2.22
5	.0092	.0356	3.14
6	.0105	.0461	4.13
7	.0119	.0580	5.36
8	.0198	.0777	6.98
9	.0237	.1014	9.13
10	.0303	.1318	11.73
11	.0277	.1594	14.91
12	.0514	.2108	18.56
13	.0303	.2411	22.75
14	.0606	.3017	27.51
15	.0527	.3544	33.27
16	.0883	.4427	40.62
17	.0988	.5415	49.98
18	.1344	.6759	61.29
19	.1238	.7997	73.89
20	.1410	.9407	85.95
21	.0593	1.0000	95.67

表11　テストM4の$f(x)$, $F(x)$, $P(x)$

x	$f(x)$	$F(x)$	$P(x)$
0	.0013	.0013	.07
1	.0026	.0040	.24
2	.0013	.0053	.51
3	.0053	.0105	.97
4	.0119	.0224	1.69
5	.0079	.0303	2.64
6	.0119	.0422	3.93
7	.0264	.0685	5.80
8	.0277	.0962	8.45
9	.0395	.1357	12.19
10	.0632	.1989	17.24
11	.0698	.2688	23.67
12	.0804	.3491	31.47
13	.1041	.4532	40.82
14	.1225	.5758	52.11
15	.1436	.7194	65.68
16	.1779	.8972	80.15
17	.1028	1.0000	93.61

(6) テスト得点の対応づけ

5 中学校で実施されたテスト X と全国学力・学習状況調査（テスト M's）のデータに基づいて，両テストの得点の対応づけを行う．上述した方法を用いて，テスト M の得点をテスト X の得点及びテスト X の得点可能な値に丸めて対応づけた結果を表 12，その対応の様子を図 7 に示す．原理的には，テスト M，MA，MB，M1，M2，M3，M4 すべての得点をテスト X の得点に対応づけることが可能であるが，テスト M's の中ではテスト M が他のすべてを包含していること，テスト M とテスト X の得点段階数が一致していることから，テスト M の得点をテスト X の得点に対応づけたものが最良であると考えられる．

また，テスト X の得点をテスト M's の得点及びテスト M's の得点可能な値に丸めて対応づけた結果を表 13，表 14 に示し，それぞれの対応の様子を図 8，図 9 に描く．

表12　テストMの得点のテストXの得点への対応づけ

M	X	Xr
0	-0.10	0.00
1	-0.07	0.00
2	-0.07	0.00
3	-0.07	0.00
4	-0.07	0.00
5	-0.05	0.00
6	0.00	0.00
7	0.11	0.00
8	0.79	0.75
9	0.98	1.00
10	1.12	1.00
11	1.29	1.25
12	1.65	1.75
13	1.74	1.75
14	1.87	1.75
15	2.00	2.00
16	2.13	2.25
17	2.24	2.25
18	2.34	2.25
19	2.65	2.75
20	2.78	2.75
21	3.00	3.00
22	3.22	3.25
23	3.63	3.75
24	3.89	4.00
25	4.09	4.00
26	4.27	4.25
27	4.63	4.75
28	4.85	4.75
29	5.17	5.25
30	5.53	5.50
31	5.78	5.75
32	6.00	6.00
33	6.21	6.25
34	6.44	6.50
35	6.80	6.75
36	7.16	7.25
37	7.67	7.75
38	8.05	8.00
39	8.33	8.25
40	8.80	8.75
41	9.23	9.25
42	9.81	9.75
43	10.37	10.25
44	10.91	11.00

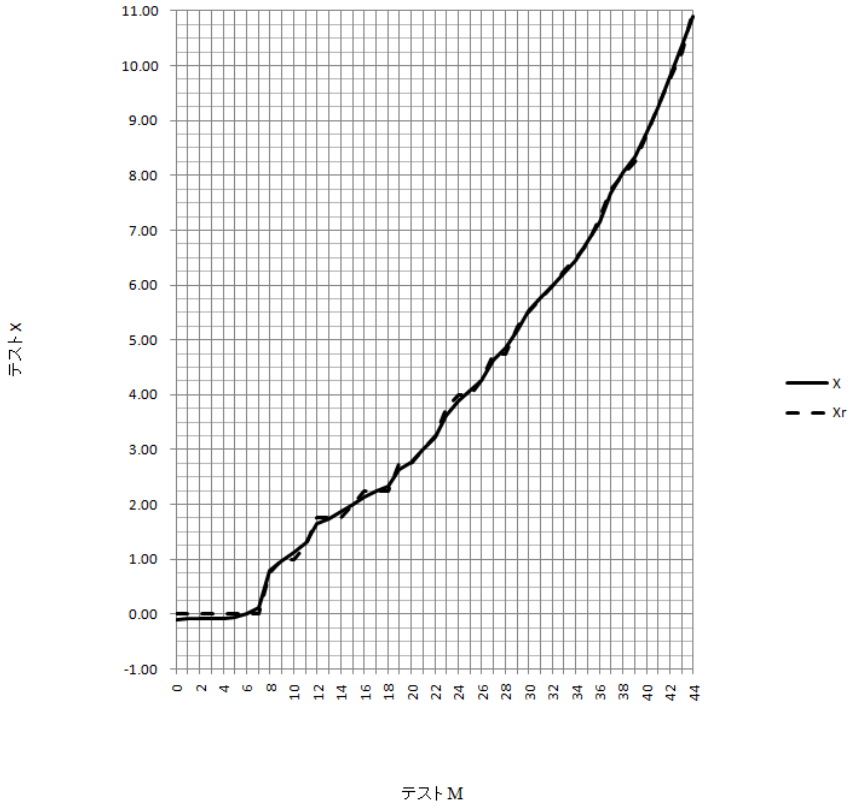


図7　テストMの得点をテストXの得点及びテストXの得点可能な値に丸めて対応づけた結果のグラフ

表13　テストXの得点のテストM'sの得点への対応づけ

X	M	MA	MB	M1	M2	M3	M4
0	5.9	5.0	-0.2	0.1	-0.3	1.6	1.2
0.25	7.2	6.0	0.1	0.5	-0.1	2.3	2.8
0.5	7.3	6.2	0.2	0.6	-0.1	2.3	2.8
0.75	7.8	6.8	0.4	0.6	0.0	2.6	3.2
1	9.3	7.9	0.7	0.7	0.3	3.0	3.7
1.25	10.8	9.0	1.0	0.8	0.5	3.3	4.0
1.5	11.6	9.5	1.2	0.9	0.6	3.5	4.2
1.75	13.0	10.4	1.7	1.0	0.8	4.3	4.8
2	15.0	12.0	2.6	1.4	1.2	5.6	6.0
2.25	17.1	13.2	3.0	1.6	1.6	6.7	6.7
2.5	18.7	14.1	3.3	1.7	1.8	7.4	7.0
2.75	19.8	14.9	3.5	1.8	2.0	7.9	7.4
3	21.0	15.9	4.0	1.9	2.4	8.6	7.9
3.25	22.2	16.9	4.5	2.1	2.7	9.3	8.5
3.5	22.8	17.6	4.8	2.3	2.8	9.8	8.8
3.75	23.4	18.1	5.1	2.4	3.0	10.2	9.2
4	24.5	18.9	5.6	2.6	3.3	10.9	9.6
4.25	25.9	19.7	5.9	2.6	3.6	11.6	10.0
4.5	26.8	20.1	6.1	2.7	3.7	11.9	10.2
4.75	27.6	20.5	6.3	2.7	3.8	12.2	10.4
5	28.6	21.1	6.6	2.8	4.1	12.8	10.8
5.25	29.2	22.0	6.9	2.9	4.3	13.6	11.2
5.5	29.9	22.7	7.2	3.0	4.6	14.0	11.5
5.75	30.8	23.3	7.5	3.1	4.8	14.4	11.9
6	32.0	24.3	7.8	3.3	5.0	15.1	12.3
6.25	33.2	25.2	8.1	3.4	5.3	15.7	12.8
6.5	34.1	25.8	8.3	3.5	5.5	16.1	13.1
6.75	34.8	26.4	8.5	3.6	5.7	16.5	13.4
7	35.6	26.8	8.8	3.6	5.9	16.8	13.7
7.25	36.3	27.2	9.0	3.7	6.1	17.2	13.9
7.5	36.7	27.6	9.2	3.7	6.2	17.5	14.2
7.75	37.2	27.9	9.3	3.8	6.4	17.7	14.4
8	37.9	28.4	9.6	3.8	6.6	18.0	14.7
8.25	38.7	29.0	9.8	3.9	6.8	18.4	15.1
8.5	39.5	29.5	10.0	4.0	7.0	18.7	15.4
8.75	39.9	29.9	10.2	4.1	7.2	19.0	15.6
9	40.4	30.4	10.4	4.1	7.4	19.3	15.8
9.25	41.1	30.8	10.6	4.2	7.6	19.7	16.1
9.5	41.6	31.1	10.8	4.3	7.8	19.9	16.3
9.75	41.9	31.4	11.0	4.3	7.9	20.2	16.5
10	42.2	31.7	11.1	4.4	8.1	20.4	16.8
10.25	42.6	32.2	11.3	4.4	8.3	20.7	17.1
10.5	43.2	32.5	11.4	4.5	8.4	21.1	17.3
10.75	43.5	33.0	11.4	4.5	8.4	21.3	17.4
11	44.2	33.4	11.5	4.5	8.5	21.4	17.5

表14　テストXの得点のテストM'sの得点のうちの可能な値への対応づけ

X	Mr	Mar	MBr	M1r	M2r	M3r	M4r
0	6	5	0	0	0	2	1
0.25	7	6	0	1	0	2	3
0.5	7	6	0	1	0	2	3
0.75	8	7	0	1	0	3	3
1	9	8	1	1	0	3	4
1.25	11	9	1	1	0	3	4
1.5	12	10	1	1	1	4	4
1.75	13	10	2	1	1	4	5
2	15	12	3	1	1	6	6
2.25	17	13	3	2	2	7	7
2.5	19	14	3	2	2	7	7
2.75	20	15	4	2	2	8	7
3	21	16	4	2	2	9	8
3.25	22	17	4	2	3	9	9
3.5	23	18	5	2	3	10	9
3.75	23	18	5	2	3	10	9
4	25	19	6	3	3	11	10
4.25	26	20	6	3	4	12	10
4.5	27	20	6	3	4	12	10
4.75	28	20	6	3	4	12	10
5	29	21	7	3	4	13	11
5.25	29	22	7	3	4	14	11
5.5	30	23	7	3	5	14	12
5.75	31	23	8	3	5	14	12
6	32	24	8	3	5	15	12
6.25	33	25	8	3	5	16	13
6.5	34	26	8	4	6	16	13
6.75	35	26	9	4	6	16	13
7	36	27	9	4	6	17	14
7.25	36	27	9	4	6	17	14
7.5	37	28	9	4	6	17	14
7.75	37	28	9	4	6	18	14
8	38	28	10	4	7	18	15
8.25	39	29	10	4	7	18	15
8.5	39	29	10	4	7	19	15
8.75	40	30	10	4	7	19	16
9	40	30	10	4	7	19	16
9.25	41	31	11	4	8	20	16
9.5	42	31	11	4	8	20	16
9.75	42	31	11	4	8	20	16
10	42	32	11	4	8	20	17
10.25	43	32	11	4	8	21	17
10.5	43	33	11	4	8	21	17
10.75	44	33	11	4	8	21	17
11	44	33	11	4	8	21	17

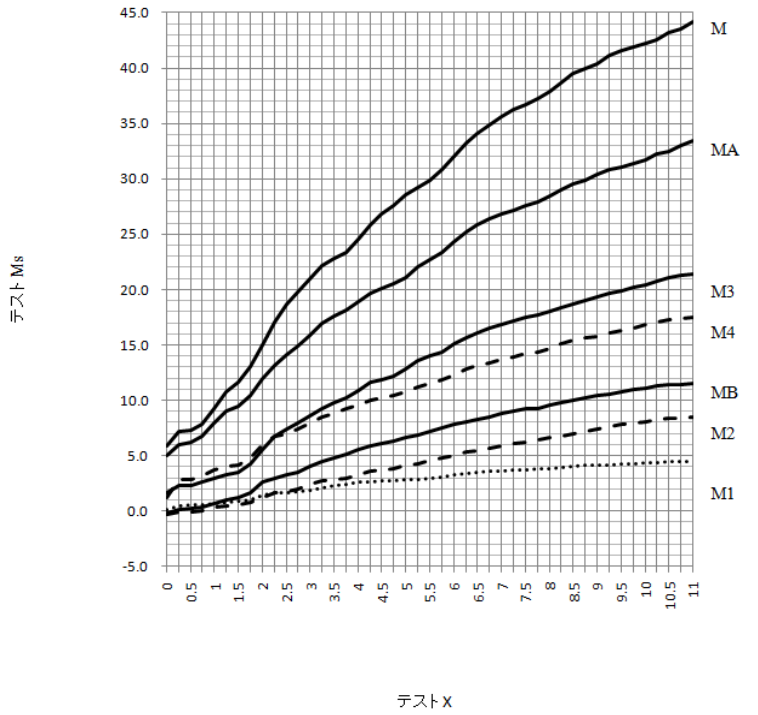


図8　テストXの得点をテストM'sの得点に対応づけたグラフ

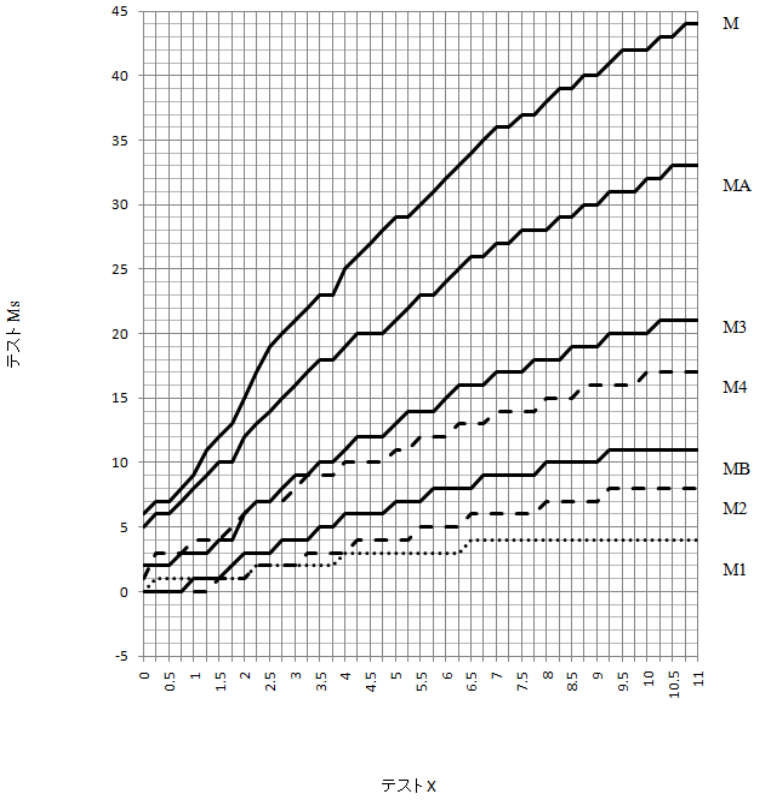


図9　テストXの得点をテストM'sの得点のうちの可能な値に対応づけたグラフ

(7) テスト得点分布の経年比較

表 12 にある得点の対応に基づいて，平成 21 年度におけるテスト M の得点から，平成 21 年度におけるテスト X の得点を推定する．また，表 13，表 14 に基づいて，平成 18 年度におけるテスト X の得点から，平成 18 年度におけるテスト M's の得点を推定する．これら推定された得点データを用いて，テスト X とテスト M's の得点分布について，平成 18 年度と平成 21 年度との経年比較を行う．

各得点の記述統計量を表 15，平成 21 年度におけるテスト M's 間の層関係数を表 16 に示す．表 15 における"r"は，得点可能な値に丸められた場合を示している．

効果量とは，平均値の差を標準偏差（standard deviation，SD）で割った値であり，相対的な平均値の変化量を表すものである．本研究では，平成 21 年度の平均値から平成 18 年度の平均値を引き，テスト X は平成 18 年度の SD，テスト M's は平成 21 年度の SD で割った値を用いている．

表15　各テストの得点分布の基本統計量，効果量，分散比，α係数

	人数	平均	SD	効果量	分散比	最小値	Q1	中央値	Q3	最大値	α係数
X 18	3892	7.27	2.30	-	-	0.00	6.00	7.50	9.00	11.00	0.70
X 21r	19498	7.15	2.36	-0.05	1.06	0.00	5.75	7.75	8.75	11.00	–
X 21	19498	7.16	2.38	-0.05	1.07	-0.10	5.78	7.67	8.80	10.91	–
M 18	3892	35.06	7.31	-0.05	1.15	5.9	32.0	36.7	40.4	44.2	-
M 18r	3892	35.11	7.26	-0.06	1.16	6	32	37	40	44	-
M 21	19498	34.65	7.83	-	-	0	31	37	40	44	0.91
MA 18	3892	26.40	5.37	-0.05	1.13	5.0	24.3	27.6	30.4	33.4	-
MA 18r	3892	26.35	5.34	-0.04	1.15	5	24	28	30	33	-
MA 21	19498	26.13	5.72	-	-	0	24	28	30	33	0.88
MB 18	3892	8.71	2.31	-0.08	1.17	-0.2	7.8	9.2	10.4	11.5	-
MB 18r	3892	8.76	2.22	-0.10	1.26	0	8	9	10	11	-
MB 21	19498	8.52	2.49	-	-	0	7	9	10	11	0.78
M1 18	3892	3.56	0.80	-0.09	0.99	0.1	3.3	3.7	4.1	4.5	-
M1 18r	3892	3.56	0.75	-0.09	1.13	0	3	4	4	4	-
M1 21	19498	3.49	0.79	-	-	0	3	4	4	4	0.43
M2 18	3892	5.96	1.88	-0.06	1.05	-0.3	5.0	6.2	7.4	8.5	-
M2 18r	3892	5.95	1.90	-0.05	1.03	0	5	6	7	8	-
M2 21	19498	5.84	1.93	-	-	0	5	6	7	8	0.68
M3 18	3892	16.51	3.94	-0.05	1.14	1.6	15.1	17.5	19.3	21.4	-
M3 18r	3892	16.50	3.84	-0.05	1.20	2	15	17	19	21	-
M3 21	19498	16.29	4.20	-	-	0	14	18	19	21	0.85
M4 18	3892	13.64	2.97	-0.08	1.18	1.2	12.3	14.2	15.8	17.5	-
M4 18r	3892	13.65	2.91	-0.08	1.23	1	12	14	16	17	-
M4 21	19498	13.39	3.22	-	-	0	12	14	16	17	0.80

表16　平成21年度のテストM's間の相関係数

	M 21	MA 21	MB 21	M1 21	M2 21	M3 21	M4 21
M 21	1						
MA 21	0.98	1					
MB 21	0.89	0.78	1				
M1 21	0.62	0.63	0.50	1			
M2 21	0.86	0.80	0.87	0.48	1		
M3 21	0.95	0.90	0.94	0.54	0.85	1	
M4 21	0.92	0.94	0.73	0.49	0.73	0.79	1

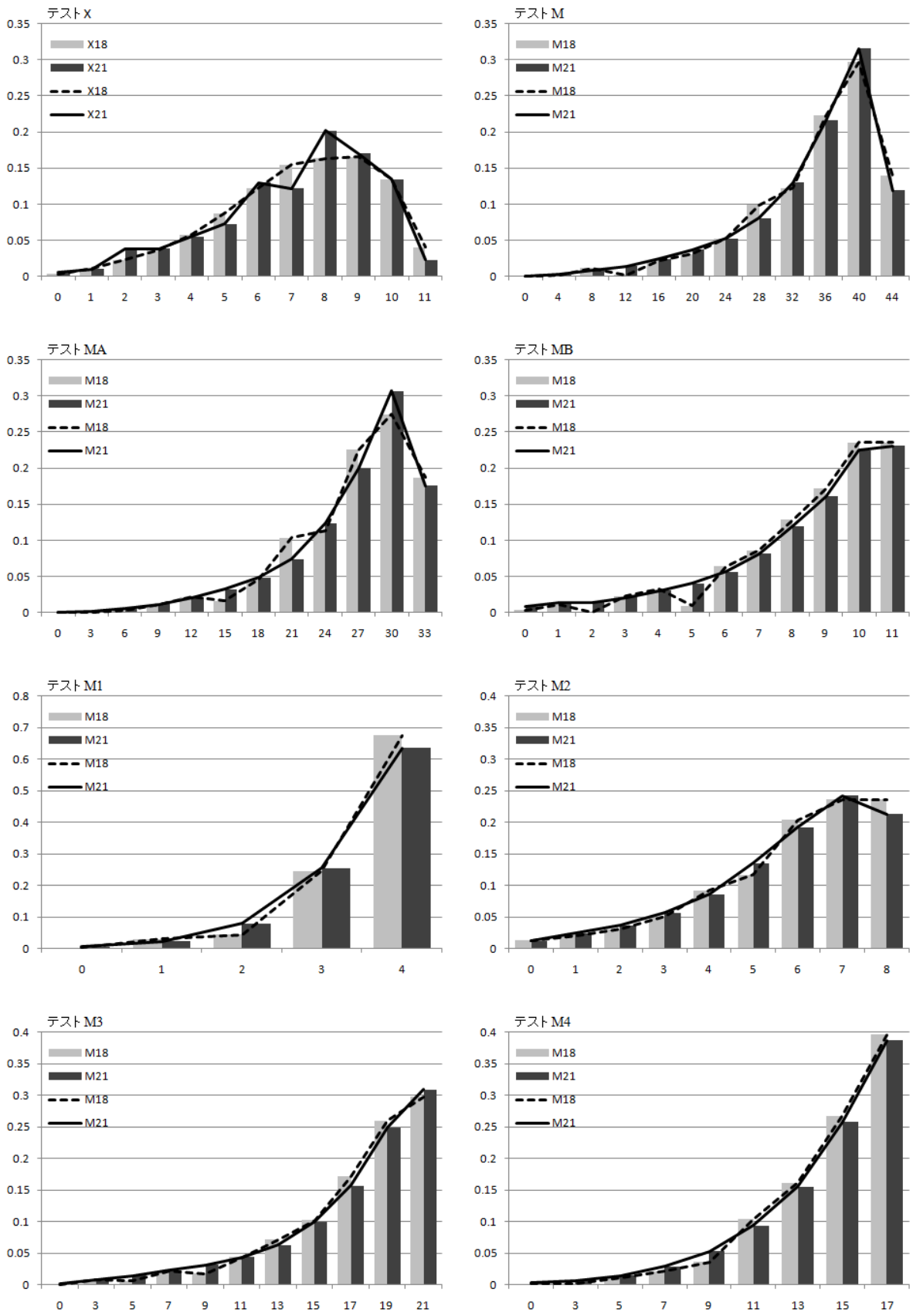


図10　各テストの得点分布の経年比較

このように計算するのは，分母の SD は実際のデータによるものであることから，実際のテスト得点の分布において，平均値が平成 18 年度と平成 21 年度の間でどれだけ変化したかを相対的に捉えられるためである．なお，過去に比べ現在の平均値が高ければプラス，低ければマイナスの値になる．

また，分散比とは，平成 21 年度の分散を平成 18 年度の分散で除した値であり，得点の分布が広がったのか狭まったのかを表すものである．過去に比べ現在の得点分布が広ければ 1 より大きい値になり，過去に比べ現在の得点分布が狭ければ 1 より小さい値になる．なお，「標準偏差（standard deviation, SD）」は分散の正の平方根であるから，もし SD の比を考えるなら，単純に分散比の平方根を計算すればよい．

Q1,Q3 は第 1 四分位数（25 パーセンタイル），第 3 四分位数（75 パーセンタイル）であり，それぞれ下位 4 分の 1，下位 4 分の 3 のところに位置する得点を表す．なお，Q2（第 2 四分位数，50 パーセンタイル）は中央値である．

得点分布のグラフを図 10 に示す．平成 18 年度のテスト M's と平成 21 年度のテスト X のグラフは，対応づけ得点そのものではなく，得点可能な値に丸めた値を用いて作成している．対応づけ得点の分布関数から得点可能な値に関するパーセンタイル順位を求めてグラフを作成することも可能であるが，得点が等間隔ではなく補完推定を数回繰り返す必要がある．これに対し，得点可能な値に丸めた値を用れば簡便にグラフを作成でき，分布の様子も十分に把握できる．

図 10 における各階級の相対度数は，当該階級の最大値と最小値のパーセンタイル順位の差の値を求めることにより算出している．テスト MB,M1,M2 は 1 点刻みで階級を構成しているので，各得点の相対度数をそのまま用いることができる．テスト X,M,MA,M3,M4 は得点を群にして階級を構成している．例えば，テスト X のグラフは 0.25 点刻みではなく 1 点刻みで階級を構成している．このようにした理由は，図 5 のテスト X の相対度数分布が煙突状であり，得点カテゴリを合併して表示したほうが分布の概観を捉えやすいと判断されるためである．階級 7 の最大値は 7.5 点，最小値は 6.5 点であり，平成 21 年度の階級 7 の相対度数 0.122 は，7.5 点のパーセンタイル順位 47.1 から 6.5 点のパーセンタイル順位 34.9 を引いた値を 100 で割って算出している．

表 15 及び図 10 のテスト X とテスト M の結果を見ると，平成 18 年度と平成 21 年度の得点分布の比較において，効果量（相対的な平均値の差）はほぼゼロであり差がほとんどないこと，分散比の大きさも 1 に近くほとんど変わらないこと，四分位数の変化も 1 点以内（テスト X においては 0.25 点以内）であるから，平成 18 年度と平成 21 年度とで，自治体テスト及び全国学力・学習状況調査の得点分布にはあまり変化はないと考えられる．

強いて言えば，効果量が負の値であるから平均点が上昇しているとは言えないが，大きく低下しているわけでもなく，また，分散比が 1 より大きいことから得点の分布が狭くなっているとは言えないが，大きく広がっているというほどでもない，すなわち，学力が大きく低下しているわけではなく，格差が大きく拡大しているというほどでもない，といったところである．

ただし，問題 A「知識」と問題 B「活用」の結果を比較すると，問題 A に比べ問題 B のほうが効果量の絶対値および分散比の値が大きいことから，もし，平成 18 年度と平成 21 年度を比較して経年変化があるとすれば，「知識」よりも「活用」においてその傾向が表れていると言うことはできるであろう．

領域別で見た場合は，もし経年変化があるとすれば，領域 2「書くこと」や領域 3「読むこと」よりも，領域 1「話すこと・聞くこと」や領域 4「言語事項」においてその傾向が見られるということになる．ただし，領域 1 については，設問数が少なく信頼性（α係数）も低いので，解釈は慎重に行うべきである．

これらは一見矛盾した結果に見える．何故なら，領域 1 及び領域 4 は，問題 A に含まれる設問のみで構成されているが，領域 2 及び領域 3 は問題 A と問題 B に含まれる設問から構成されており，もし問題 A よりも問題 B において経年変化が見られるとすれば，領域 1 や領域 4 よりも，領域 2 や領域 3 において経年変化が表れやすいと考えられるからである．

しかし，効果量は相対的な平均値の変化を表すものであり，問題 A の設問数（33 問）のほうが問題 B の設問数（11 問）よりも 3 倍も大きいことを考えれば，問題 A または問題 B から抽出される集合の部分や大きさの違いによっては，問題 A による効果のほうが前面に出てくることは十分あり得ることであると理解できる．

表17　領域2及び領域3の得点に関する基本統計量，効果量，分散比，α係数

	人数	平均	SD	効果量	分散比	最小値	Q1	中央値	Q3	最大値	α係数
M2A 18	3892	2.12	0.96	-0.11	0.92	-0.5	1.6	2.2	2.9	3.5	-
M2A 18r	3892	2.11	0.90	-0.10	1.06	0	2	2	3	3	-
M2A 21	19498	2.02	0.93	-	-	0	1	2	3	3	0.40
M2B 18	3892	3.88	1.28	-0.05	1.04	-0.4	3.3	4.1	4.8	5.5	-
M2B 18r	3892	3.84	1.25	-0.01	1.09	0	3	4	5	5	-
M2B 21	19498	3.82	1.31	-	-	0	3	4	5	5	0.61
M3A 18	3892	7.86	1.97	-0.04	1.08	0.3	7.2	8.4	9.2	10.5	-
M3A 18r	3892	7.86	1.92	-0.04	1.13	0	7	8	9	10	-
M3A 21	19498	7.77	2.04	-	-	0	7	8	9	10	0.72
M3B 18	3892	8.71	2.31	-0.08	1.17	-0.2	7.8	9.2	10.4	11.5	-
M3B 18r	3892	8.76	2.22	-0.10	1.26	0	8	9	10	11	-
M3B 21	19498	8.52	2.49	-	-	0	7	9	10	11	0.78

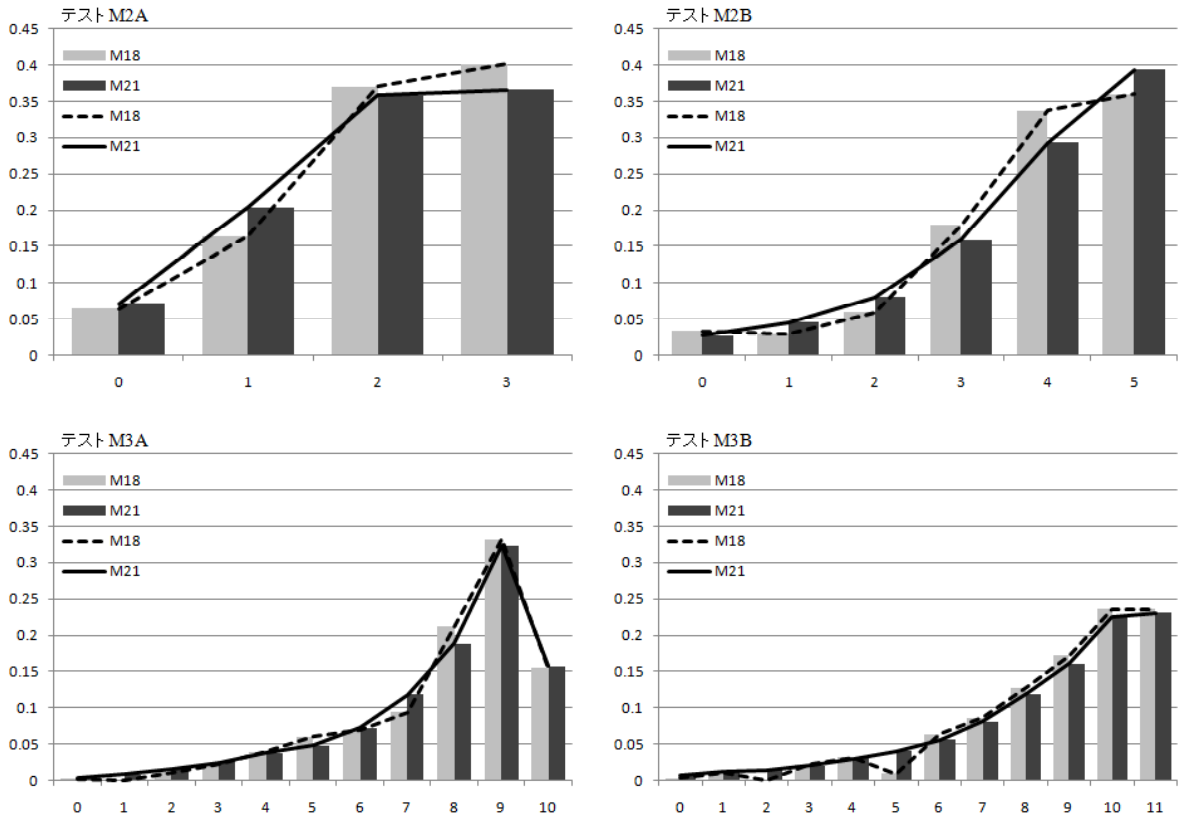


図11　領域2及び領域3の得点分布の経年比較

領域 2 と領域 3 について，問題 A に含まれる設問と問題 B に含まれる設問とを分割した場合の得点の記述統計量を表 17，得点分布のグラフを図 11 に示す．領域 2 に含まれる問題 A の設問は 3 設問，問題 B の設問は 5 設問，領域 3 に含まれる問題 A の設問は 10 設問，問題 B の設問は 11 設問である．それぞれのテストを M2A，M2B，M3A，M3B と表記する．推定の方法はこれまでと同様である．

表 17 を見ると，領域 2「書くこと」においては，問題 B「活用」よりも問題 A「知識」における変化のほうが相対的に大きく，領域 3「読むこと」においては，反対に問題 A「知識」よりも問題 B「活用」における変化のほうが相対的に大きいことがわかる．ただし，領域 2 については，設問数が少なく信頼性（α 係数）も低いので，解釈は慎重に行うべきである．

6. データ収集デザインの拡張

本節では，テスト数，版数，時点数，地域数が増えた場合にも，同様の得点の対応づけを行うようにするデータ収集デザインについて述べる．

(1) 両地域において経年比較を行う場合

図 1 のデータ収集デザインでは，地域 B のデータは得点の対応づけに用いられるだけであった．もし地域 B においても独自のテストが実施されていて，得点分布の経年比較を行うならば，図 12 に示すようなデザインを組めばよい．

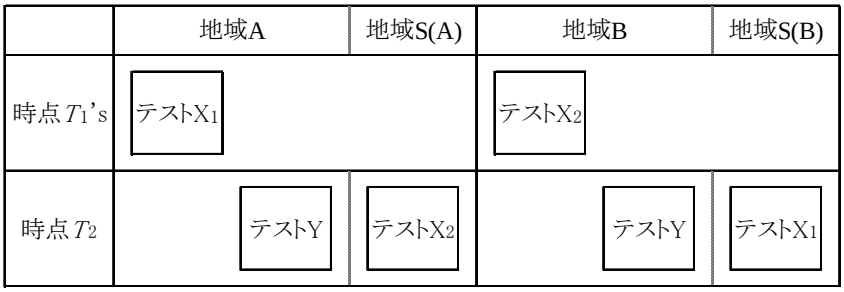


図12　地域A，地域Bにおけるテスト得点分布の経年比較を行うデータ収集デザイン

図 12 において，テスト X_1,X_2,Y はいずれも同一教科のテストである必要がある．以降のデザインでも同様である．もし複数教科を扱う場合には，それぞれの教科について，データ収集デザインを組むことになる．

地域 A においてテスト X_1 を実施する時期と，地域 B においてテスト X_2 を実施する時期は必ずしも同じでなくてよい．このことを図では，T_1's と表記することによって表している．また，時点 T_2 の地域 A でのテスト X_2 および地域 B でのテスト X_1 の実施は，必ずしも当該地域全体でなされる必要はなく，ある程度の規模の特定可能な集団であればよい（もちろん，地域全体でもよい）．このことを図では，S(.)という記号を用いて表している．

このようなデザインを組むことにより，地域 S(A)のデータからテスト X_2 とテスト Y の対応づけが得られるので，それを地域 B のデータに当てはめて，地域 B における時点 T_1 と T_2 のテスト X_2 およびテスト Y の得点分布の経年比較を行うことが可能となる．

(2) 自治体テストが複数年度実施される場合

次に，地域 A または地域 B において独自に実施されるテストが複数個ある場合の経年比較の方法を考える．この場合は，図 13 に示すようなデザインを組むことにより，地域 A における時点 T_1, T_2, T_3 のテスト X_1, テスト X_3 およびテスト Y の得点分布, また, 地域 B における時点 T_1, T_2, T_3 のテスト X_2，テスト X_4 およびテスト Y の得点分布の経年比較を行うことが可能となる．

テスト X_1 とテスト X_2，また，テスト X_3 とテスト X_4 を実施する時期は必ずしも同じでなくてよい．また，地域 A の部分集団において実施されるテスト X_2 とテスト X_4 は同一集団を対象とする必要はなく，同様に，地域 B の部分集団において実施されるテスト X_1 とテスト X_3 も同一集団を対象とする必要はない．さらに，時点 T_1，T_2，T_3 の順番もデザイン上の制約はなく（実施上の制約はあり得る），例えば T_2 が最後でも構わない．

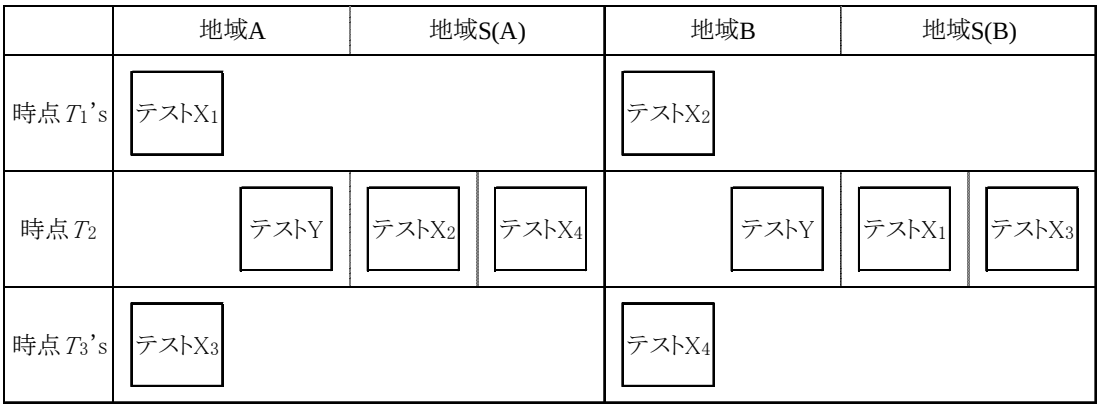


図13　地域A，地域Bにおける複数個のテスト得点分布の経年比較を行うデータ収集デザイン

(3)　全国学力・学習状況調査が複数年度実施される場合

今度は，共通に実施されるテストが複数の時点でそれぞれある場合を考える．それらをテスト Y_1，テスト Y_2 とする．もし地域 A と地域 B でそれぞれ独自に行われるテスト（テスト X_1，テスト X_2）が，テスト Y_1 と同じ時点 T_1 で実施されていれば，図 14 に示すようなデザインを組むことにより，地域 A においてはテスト X_1，テスト Y_1，テスト Y_2，地域 B においてはテスト X_2，テスト Y_1，テスト Y_2 の得点分布の経年比較を行うことが可能となる．

ここで，地域 A における時点 T_1 と T_2 のテスト X_1 とテスト Y_2 の経年比較は，図 1 や図 12 のデザインと同様になし得るが，テスト Y_1 については，時点 T_1 におけるテスト X_1 とテスト Y_1 の得点分布の対応づけを行っておき，その対応関係を，時点 T_2 の推定されたテスト X_1 の得点分布に当てはめてテスト Y_1 の得点分布を推定するというように，対応づけを 2 回行って推定する．地域 B の場合も同様である．

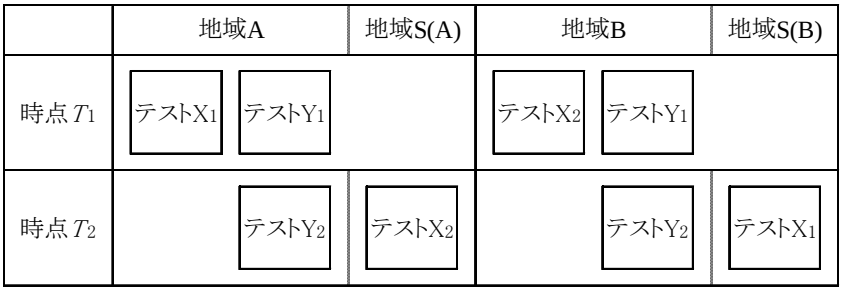


図14　共通テストが複数回ある場合のテスト得点分布の経年比較を行うデータ収集デザイン

(4)　全国学力・学習状況調査が単年度に複数版実施される場合

共通に実施されるテストが複数の時点でそれぞれあるのではなく，同一時点で複数の共通テストがある場合は，図 15 に示すようなデザインを組めばよい．

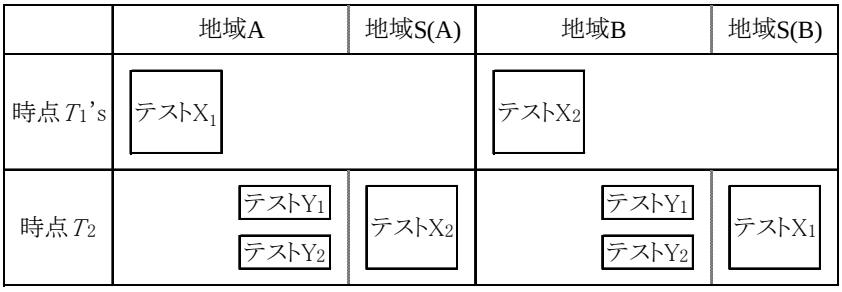


図15　共通テストが複数個ある場合のテスト得点分布の経年比較を行うデータ収集デザイン

地域 A においてはテスト X_1，テスト Y_1，テスト Y_2，地域 B においてはテスト X_2，テスト Y_1，テスト Y_2 の得点分布の経年比較を行うことが可能となる．この場合，テスト Y_1 とテスト Y_2 は同一教科のテストであるが，異なる項目で構成することができるため（もちろん，共通項目があってもよい），1 回の実施でより広い範囲の学力を調査することができるという利点を有する．実際，全米学力調査（NAEP）など多くの大規模テストで，このような方法が取られている（荒井・倉元, 2008; 村木, 2006 など）．

(5) 地域が多数ある場合

ここまでのデザインは，2 つの地域でテストが実施されることを想定して作成されたものである．しかし，例えば図 12 を螺旋状に展開して，3 つの地域を扱うものに拡張することもできる．その様子を図 16 に示す．図 16 において，テスト X_1，テスト X_2，テスト X_3 を実施する時期は必ずしも同じでなくてよい．

図 16 のようなデザインを組むことにより，地域 A における時点 T_1,T_2 のテスト X_1 とテスト Y の得点分布，地域 B における時点 T_1,T_2 のテスト X_2 とテスト Y の得点分布，地域 C における時点 T_1,T_2 のテスト X_3 とテスト Y の得点分布の経年比較を行うことがそれぞれ可能となる．

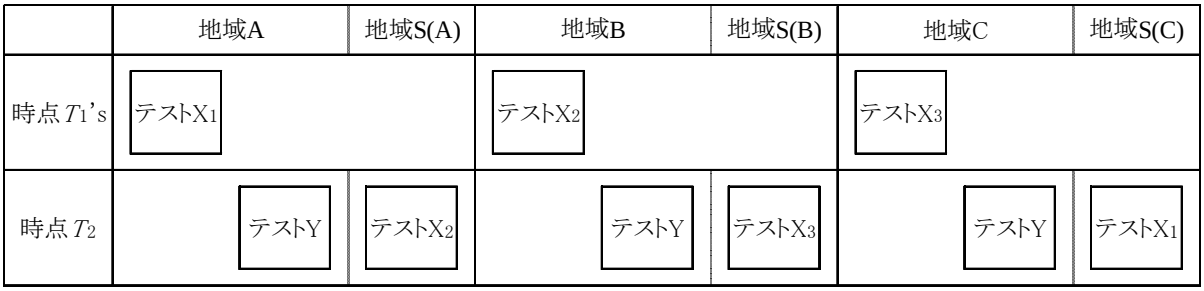


図16　3つ以上の地域がある場合のテスト得点分布の経年比較を行うデータ収集デザイン

(6) 拡張デザインと実際との対応

上記に示したデータ収集デザインを組み合わせることにより，さらに複雑なデザインを組むこともちろん可能であるが，要点は，「ある地域 A において時点 T_1 に実施されたテスト X があり，地域 A と地域 B において時点 T_2 にテスト Y が実施され，また地域 B（S(B)でもよい）においては時点 T_2 にテスト X も実施されていて，その集団が特定できるという構造があれば，地域 A における時点 T_1 と T_2 のテスト X とテスト Y の得点分布の経年比較を行うことができる」ということである．

一般化して述べてきたことを具体的に考えてみよう．「地域」を自治体とすれば，複数の自治体が協同し，あらかじめ計画されたデータ収集デザインに基づいて，独自のテストを互いに交換し，時期や対象集団を調整してテストが実施されれば，それぞれの自治体において，テスト得点の経年変化を客観的に捉えることが可能となり，それを教育実践，教育施策等に活かすことができることになる．

なお，テスト問題が公開されてしてしまうことの問題点を冒頭で述べたが，協力しあう自治体が近接しておらず，得点の対応づけのためのデータを収集する集団が事前には分からないようにしておけば,試験対策が取られてしまうことによる問題はある程度回避されるものと考えられる．経年比較を行う時点の間隔が広ければ，さらに問題は低減するであろう．

自治体間の協力関係が一般には公表されていないほうがデータ収集上は望ましいが，それは難しい場合も多いかもしれない．そのような場合には，複数の自治体が協同し，互いに協同相手を限定せず，年度によって相手を変える，協同相手は当日まで非公開とする，などの対応も考えられる．

7. まとめ

本研究で行ったことを簡潔にまとめると以下のようになる.

本研究では，全国学力・学習状況調査と，自治体等により作成・実施されたテストと組み合わせることにより，学力（テストの得点分布）の経年比較を行う方法を提案し，その有効性を実際のデータを用いて提示することを目的とした.

前半の学力の経年比較を行う方法の提案については，図 1 で示されるようなデータ収集デザインに基づいてテストが実施されれば，教育測定学の領域で研究されている対応づけの手法を用いて，全国学力・学習状況調査および自治体等によって作成・実施されたテストの得点を対応づけることができ，得点分布の経年比較を行うことが可能であることを示した.

後半の適用例については，自治体テストと全国学力・学習状況調査のデータを用いて，平成 18 年度と平成 21 年度の中学校 3 年生国語テストの得点分布の経年比較を行った. その結果, 平成 18 年度と平成 21 年度とで，自治体テスト及び全国学力・学習状況調査の得点分布にはあまり変化はないが，もし変化があるとすれば,「知識」よりも「活用」において，また,「書くこと」や「読むこと」よりも「言語事項」において，その傾向が見られることなどが見出された.

また，データ収集デザインを拡張し，テスト数，版数，時点数，地域数を増やした場合などについても言及した.

次に，今後の課題について述べる.

本研究で提案した方法は，2 つのテストのデータを使い，一方から他方の得点を推測することにより得点分布の経年比較を行うというものである．よって，そこには標本誤差や推定誤差の影響が含まれている．適用例において微小な効果量の値や分散の変動に対して過剰な解釈を行わなかったのは，この影響を考慮したことにもよる.

しかし，この微小と見なした差異が，教育現場では大きな意味を持つかもしれないし，反対に，もっと大きな差異があっても教育現場ではさほど影響ないかもしれない．このことを検証することをはじめ，テストを実施することが教育実践，教育施策等にさらに有効となるためには，教育実践に携わる者と，テスト及びテストデータを分析する研究者とが協同して，より多くの研究を積み重ねていく必要があると考えられる.

また，本研究で用いた得点の対応づけの方法は，等パーセンタイル法という古典的な手法であったが，先にも述べたように，項目応答理論（IRT）を用いれば，個々の受検者の能力値をテストに依存しない共通尺度上で推定できること，項目困難度を等化することにより各テストに含まれる項目の難しさを相互比較できるようになること，項目情報量やテスト情報量を算出することによりどの層の受検者の学力をよく識別するテストであるかを確認できることなど，より多くの情報を得ることができる．項目応答理論が適用できる場面においては，同手法による得点の等化を行うことも検討する必要があるであろう.

さらに，もし能力分布の比較が目的ならば，個々の受検者の能力値を推定しないでも分布の推定が行える Plausible Value Technology（Mislevy, et al., 1992）を適用することも考えられる．この手法は，標本誤差による影響を少なくする方法として有効であり，全米学力調査（NAEP）でも利用されている．今後はこうした手法を用いるだけでなく，さらに研究を行い，わが国の実情にあったテスト理論，テスティング技術を開発していく必要があるであろう.

文献

荒井克弘・倉元直樹(編著)（2008）全国学力調査 日米比較研究. 金子書房.

Brennan, R.L., & Kolen, M.J.（1987）Some practical issues in equating. Applied Psychological Measurement, 11, 279-290.

Downing, S.M., & Haladyna, T.M., eds.（2006）Handbook of test development, Lawrence Erlbaum Associates, Inc.（池田央(監訳)(2008）テスト作成ハンドブックー発達した最新技術と考え方による公平妥当なテスト作成・実施・利用のすべて，教育測定研究所）

Holland, P.W., & Dorans, N.J.（2008）Linking and Equating. In: Brennan, R.L.. ed. Educational measurement, 4th ed. American Council on Education and Prager Publishers.

池田央（1994）現代テスト理論. 朝倉書店.

木村拓也（2006）戦後日本において「テストの専門家」とは一体誰であったのか？ー戦後日本における学力調査一覧と「大規模学力テスト」の関係者一覧. 教育情報学研究, 4, 67-100.

Kolen,M.J.（1991）Smoothing methods for estimating test score distributions. Journnal of Educational Measurement, 3(1), 97-104.

Kolen,M.J. & Brennan, R.L.（2004）Test equating, scaling and linking: Methods and practices, 2nd ed. Springer.

Lord, F.M.（1982）The standard error of equipercentile equating. Journal of Educational Statistics, 7, 165-174.

前川眞一（1999）得点調整の方法について． 柳井晴夫・前川眞一編：大学入試データの解析ー理論と応用. 現代数学社, pp.88-109.

Mislevy,R.J., Beaton,A.E.Kaplan,B., & Sheehan,K.M.（1992）Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 133--161.

文部科学省・国立教育政策研究所（2007a）平成 19 年度全国学力・学習状況調査【小学校】報告書.

文部科学省・国立教育政策研究所（2007b）平成 19 年度全国学力・学習状況調査【中学校】報告書.

文部科学省・国立教育政策研究所（2008a）平成 20 年度全国学力・学習状況調査【小学校】報告書.

文部科学省・国立教育政策研究所（2008b）平成 20 年度全国学力・学習状況調査【中学校】報告書.

文部科学省・国立教育政策研究所（2009a）平成 21 年度全国学力・学習状況調査【小学校】調査結果概要.

文部科学省・国立教育政策研究所（2009b）平成 21 年度全国学力・学習状況調査【中学校】調査結果概要.

村上隆（2008）Q&A で知る統計データ解析 DOs and DON'Ts［第 2 版］. サイエンス社.

村木英治（2006）全米学力調査(NAEP)概説－テストデザインと統計手法について. 東京大学大学院教育学研究科教育測定・カリキュラム開発講座 2005 年度研究活動報告書(2) Sokutei Report, 3, 51-66.

斉田智里（2003）高校入学時の英語能力値の年次推移－項目応答理論を用いた県規模英語学力テストの共通尺度化. STEP BULLETIN(日本英語検定協会), 15, 12-24.

芝祐順・南風原朝和（1990）行動科学における統計解析法. 東京大学出版会.

柴山直（2008）日本のテスト文化について. 人事試験研究, 208, 2-13.

柴山直・野口裕之（2004）「対応づけ」の理論と計算アルゴリズム. 適性試験委員会：法科大学院統一適性試験テクニカル・レポート 2004. 商事法務研究会, pp.53-72.

柳井晴夫・石井秀宗（2008）大規模学力テストと学ぶ力に関する研究をめぐって. 児童心理学の進歩, 47, 57-86.

吉村宰・荘島宏二郎・杉野直樹・野澤健・清水裕子・齋藤栄二・根岸雅史・岡部純子・Fraser, S.（2005）大学入試センター試験既出問題を利用した共通受験者計画による英語学力の経年変化の調査. 日本テスト学会誌, 1, 51-58.

平成 21 年度文部科学省企画公募委託研究
「学力調査を活用した専門的な課題分析に関する調査研究」
『地域におけるデータ等を補完的に用いた調査分析手法の調査研究』　報告書

2010 年 4 月 1 日　発行

発行者　　石井　秀宗
〒 464-8601　名古屋市千種区不老町
名古屋大学大学院教育発達科学研究科