UDC 371.279.1

多重指標モニタリングによる得点調整手続きの試案

荘島宏二郎*
石塚　智一*
橋本　貴充*
大津　起夫*
前川　眞一**

要　約

平成 9～17 年度の大学入試センター試験「地理歴史」教科のデータの分析結果から，現行の平均点差による得点調整手続きについての利点と限界点を整理した．その結果，平均点によって分布の違いを検出するのでは不十分であることがわかった．したがって，本研究では，分布の違いをパーセンタイル得点差や絶対平均偏差，偏差平方平均平方根，完全情報最尤法による平均点差などの複数の指標によって監視するという多重指標モニタリングという得点調整手続きを提案した．

キーワード：得点調整，「地理歴史」，完全情報最尤法，ランダム欠測

1　はじめに

大学入試センター試験において利用されている得点調整の手続きは，「原則として，20 点以上の平均点差が生じ，これが試験問題の難易差に基づくものと認められる場合には，得点調整を行う（大学入試センター，2006）」というものであり，得点調整を行うかどうかの判断は得点調整判定委員会が行っている．得点調整の仕方は，分位点差縮小法（reduced percentile method, RPM；前川，2003）を用いて平均点差を 15 点までに縮小する．以上の手続きを単純に図示したものが図 1 である．具体的には，得点調整対象科目は，平成 19 年度実施教科科目で言えば，地理歴史の「世界史 B」「日本史 B」「地理 B」，公民の「現代社会」「倫理」「政治・経済」，理科の「物理 I」「化学 I」「生物 I」「地学 I」である．

実際には，平均点だけで得点調整の要否を判断することが難しい．平均は，分布の位置を示す統計量である．したがって，歪みの少ない単峰分布のとき，平均が得点調整を行う上で良い指標であるためには，比較する分布の形状が類似している必要がある．付録 A は，平成 9 (1997)～平成 17 (2005) 年度大学入試センター試験の「地理歴史」教科における 3 科目の得点分布である．「世界史 B」と「日本史 B」の分布の形状は比較的類似しているという印象があるが，「地理 B」の分布の形状は，それら 2 科目の分布と類似しているとは言えなさそうである．標準偏差（SD）・歪度・尖度を SPSS14.0（SPSS, 2005）によってもとめ，表 1 に示す．網掛けしている平成 10 年度は，実際に得点調整が行われた年である．付録 A から与える印象どおり，概して年度を通して「世界史 B」と「日本史 B」が扁平であり，「地理 B」の歪度が負である傾向があり，3 科目の SD・歪度・尖度が近い年

* 大学入試センター研究開発部 〒 153–8501 目黒区駒場 2–19–23
** 東京工業大学社会理工学研究科
2006 年 11 月 20 日　受理

図 1 現行の得点調整手続き

度はあまり見かけられない印象である．

平成 12 年度に着目しよう．表 1 より，平成 12 年度の地歴 3 科目の平均点は他の年度に比べて近いといってよいが，SD・歪度・尖度が大きく異なる．しかしながら，図 A–4（付録 A）から明らかなとおり，分布が扁平であるため，高得点取得者層においては,「世界史 B」選択者が目立っている．実際，90 点以上をとった受験者は，ほとんど「世界史 B」選択者といってよい．このことは,「日本史 B」「地理 B」選択者にとっては不利益になっている可能性がある．

平均点も分布の違いを計量化するための 1 つの有力な統計量であることに間違いない．しかし，今の例のように，平均点のみで得点調整の要否を決定することは，ある特定の科目選択者にとって著しく不利益をこうむる可能性がある．したがって，平均点以外にも，例で挙げたような平均値の周りの高次の積率の違いを議論することは重要である．また，それらのほかにも，分布の違いを計量化していくための統計量を用いることができる．本研究の目的は，平均点以外にもさまざまな統計量を用いて，分布の違いを計量化する統計量を用いて得点調整の要否を決定するための枠組みを提案することである．

ところで，経験分布の差を議論する上で，もっとも最初に思い浮かぶのは経験分布のノンパラメトリック検定である．それらは，Kruskal-Wallis の順位和検定，Mann-Whitney の U 検定，Wald-Wolfowitz のラン検定，クロス表の独立性の χ^2 検定などである．また，柳川 (1982) や福嶋・高橋・樽磨・佐藤 (2005) のようにコルモゴロフ・スミルノフ検定を用いることもできる．これらの検定方法を用いて有意であれば，分布間に差があるとみなして得点調整を行うという手続きが考えられる．

図 2 多重指標モニタリングの概念図

しかしながら，これらの統計量は，大学入試センター試験の得点調整対象科目に適用すれば，その標本サイズの大きさゆえに例外なく棄却される統計量である．したがって，これらの検定の結果によって分布間に差があるかどうかの意思決定に用いることは非現実的である．むしろ分布の差が統計的に有意か非有意かより，分布の差がいったいどれくらいなのかをしっかりと計量する方法が好ましい．

2 多重指標モニタリング

本研究の目的は，平均点以外にもさまざまな統計量や指標を参照しながら，分布の違いを計量化し，得点調整の要否を決定するための枠組み「多重指標モニタリング（multiple indices monitoring)」を提案することである．その概念図を図示すると，図 2 のようになる．

多重指標モニタリングを実際に運用するには，図 2 からも類推できるように，①指標の選択，②それぞれの指標の基準の設定，③いくつの指標に抵触したら得点調整をするか，という大きく 3 つのパラメタを設定する必要がある．①については，従来用いてきた平均点差，また，上で例示した 2 次以上の平均値周りの積率（SD・歪度・尖度)，さらには，以下で取り上げるパーセンタイル得点差（十分位点差)，絶対平均偏差（mean absolute dif-

表 1 平成 9 (1997)～17 (2005) 年度の「地理歴史」の主要統計量

		度数	平均値	標準偏差	歪度	尖度*
平成 9 (1997)	世界史 B	116,741	68.35	18.28	−0.332	−0.699
	日本史 B	179,571	55.84	17.05	−0.097	−0.750
	地理 B	121,462	67.34	13.98	−0.333	−0.284
平成 10 (1998)	世界史 B	112,269	61.03	20.56	−0.121	−0.997
	日本史 B	167,443	56.33	15.83	0.055	−0.588
	地理 B	121,606	77.23	14.50	−0.786	0.312
平成 11 (1999)	世界史 B	102,669	64.12	19.49	−0.274	−0.888
	日本史 B	153,286	58.97	17.19	−0.080	−0.745
	地理 B	118,547	62.27	14.09	−0.147	−0.451
平成 12 (2000)	世界史 B	105,272	64.62	20.31	−0.238	−0.946
	日本史 B	149,309	61.74	19.94	−0.077	−0.893
	地理 B	120,739	58.22	13.88	−0.019	−0.352
平成 13 (2001)	世界史 B	107,381	61.67	19.73	−0.158	−0.914
	日本史 B	147,993	55.26	18.46	0.129	−0.747
	地理 B	128,097	63.55	13.40	−0.305	−0.177
平成 14 (2002)	世界史 B	109,135	59.88	20.15	−0.169	−0.897
	日本史 B	154,902	58.71	18.58	−0.015	−0.862
	地理 B	127,271	66.34	16.04	−0.472	−0.327
平成 15 (2003)	世界史 B	106,583	56.53	18.56	−0.002	−0.835
	日本史 B	157,610	63.93	18.02	−0.214	−0.787
	地理 B	127,439	54.99	14.61	−0.120	−0.480
平成 16 (2004)	世界史 B	100,486	61.47	19.37	−0.091	−0.919
	日本史 B	154,808	56.52	18.16	−0.048	−0.812
	地理 B	119,555	62.11	15.81	−0.208	−0.465
平成 17 (2005)	世界史 B	93,785	63.16	20.27	−0.107	−0.979
	日本史 B	152,106	59.27	16.43	−0.115	−0.642
	地理 B	109,827	70.22	14.93	−0.553	−0.083

*尖度は，ここより以降，0 に基準化したものを用いている．

ference, MAD)，偏差平方平均平方根（root mean square difference, RMSD)，完全情報最尤法による平均差などが候補となるだろう．また，②で述べたように，それらの指標にそれぞれ基準を設ける必要がある．さらに，③で述べたように，いくつの指標で基準を満たさないときに，得点調整を行うかということも決める必要がある．無論，最終的な判断は，従来のように，得点調整判定委員会にゆだねることになるだろう．

2.1 パーセンタイル得点差（十分位点差）

各科目のパーセンタイル得点に着目し，その得点差を比較するということが考えられる．あまりパーセンタイル得点を細かく検討することは煩雑であるので，たとえば，10%刻みのパーセンタイル得点，すなわち，十分位点差（decile score difference, DSD）などが考えられるだろう．図 3 は，平成 9 (1997)～17 (2005) 年度「地理歴史」の「世界史 B」「日本史 B」「地理 B」における DSD をプロットしたものである．それぞれ，図 3–1 は「世界史 B」と「日本史 B」，図 3–2 は「世界史 B」と「地理 B」，図 3–3 は「日本史 B」と「地理 B」の DSD を表している．また，プロット中の数字は，たとえば，9 ならば第 9 十分位点（90 パーセンタイル得点，すなわち上位 10%）同士の得点差をあらわしている．したがって，5 の系列は中央値同士の差を意味している．また，M は全体の平均点差であり，参考のために記載している．全ての折れ線が 0 付近に集中していれば，そのとき当該分布間は大差ないことを示している．

図 **3–1** 「地理歴史」の十分位点差「世界史 B」と「日本史 B」

図 **3–2** 「地理歴史」十分位点差「世界史 B」と「地理 B」

図 **3–3** 「地理歴史」十分位点差「日本史 B」と「地理 B」

図 4　平成 9 (1997)～17 (2005) 年度の MAD と RMSD

図 3 から明らかなように，仮に平均点差でほとんど差がついていなかったとしても十分位点差が際立つケースが存在していることがわかる．実際に得点調整が行われた平成 10 年度のほかに，たとえば，平成 12 年は，平均点では 5 点差もついていないが，第 9 十分位点（上位 10%）には 10～15 点もの差がついている．これは，「地理 B」選択者のうち上位 10%に入る受験者とって，上位国公立大学への受験が不利となっている．

2.2　平均絶対偏差と偏差平方平均平方根

度数分布などの一次的な統計資料が与えられれば，手続きさえ示せば追試可能であるような利便性のある指標は好ましい．このとき，平均絶対偏差（mean absolute difference, MAD）と偏差平方平均平方根（root mean square difference, RMSD）は有用である．それらは，

$$MAD = \frac{\sum_q |F_q - G_q|}{hQ} \tag{1}$$

$$RMSD = \sqrt{\frac{\sum_q (F_q - G_q)^2}{hQ}} \tag{2}$$

である．ここで F_q と G_q は，得点調整対象科目の 2 つの分布において，$q(1, \cdots, Q)$ 番目の階級の度数である．また，h はバンド幅である．それらの意味するところは，MAD は「1 点あたりの受験者の割合は，平均的にどれくらいか」を問題としており，RMSD は「一方が，もう一方を基準としたとき 1 点あたりの受験者の割合の散らばり（標準偏差)」である．どちらも 0 に近いほど，2 つの分布が近いことを示しており，0 のとき両分布は完全に一致していることになる．

なお，F と G の分布に何らかの平滑化を行うなどして，連続分布として取り扱うこともあるだろう．そのとき，(1)(2) 式は，それぞれ，

$$MAD = \frac{\int_X |F(x) - G(x)| dx}{x_U - x_L} \tag{3}$$

$$RMSD = \sqrt{\frac{\int_X \{F(x) - G(x)\}^2 dx}{x_U - x_L}} \tag{4}$$

である．ここで，F と G は連続な 2 つの分布である．また，$X = [x_L, x_U]$ は x の定義域であり，100 点満点のテストであれば [0, 100]，200 点満点のテストであれば [0, 200] である．このような MAD と RMSD を平成 9 (1997)～17 (2005) 年度にかけてプロットしたものが図 4 である．なお，棒グラフが MAD をあらわし，エラーバーが RMSD をあらわしている．

2.3　完全情報最尤法による平均点差

従来の平均点差は，たとえば平成 17 年度の「地理歴史」の 3 科目には平均点差が生じているが，言うまでもなく「世界史 B」の平均 63.2 点は 93,785 人の受験者の得点から算出されたものであり，「日本史 B」の平均 59.3 点は，「世界史 B」を選択した受験生とは別の 152,106 人の得点から算出され

表 2 「地理歴史」の科目選択によるデータ欠測

受験者＼科目	世界史 B	日本史 B	地理 B
グループ 1	$X_{1W}^{(1)}$	$X_{1J}^{(0)}$	$X_{1G}^{(0)}$
グループ 2	$X_{2W}^{(0)}$	$X_{2J}^{(1)}$	$X_{2G}^{(0)}$
グループ 3	$X_{3W}^{(0)}$	$X_{3J}^{(0)}$	$X_{3G}^{(1)}$
グループ 4	$X_{4W}^{(0)}$	$X_{4J}^{(0)}$	$X_{4G}^{(0)}$

たものである．同様に，「地理 B」の平均 70.2 点は，「世界史 B」「日本史 B」を選択した受験生とは別の 109,827 人の得点から算出されたものである．むろん，それぞれの科目選択者は，その他の「地理歴史」教科における「世界史 A」「日本史 A」「地理 A」の選択者とも別の受験生である．

このことを整理すると，表 2 にようになる．表 2 は，グループ 1 が「世界史 B」を選択した受験者集団である．また，グループ 2，3 は，それぞれ，「日本史 B」「地理 B」を選択した集団である．さらに，グループ 4 は，得点調整対象科目ではないが「世界史 A」「日本史 A」「地理 A」の選択者たちと，そもそも「地理歴史」のどの科目も受験しなかった受験生を含む．いうまでもなく平成 17 年度でいえば，グループ 1 の人数が 93,785 人であり，平均点 63.2 は，データ $X_{1W}^{(1)}$ を集計して算出されたものである．ここで，$X_{gS}^{(u)}$ は，グループ g の科目 S に対するデータをあらわしている．また，上付き文字の u はデータが観測されたならば 1，そうでないならば 0 とする 2 値変数である．なお，観測された変数と欠測された変数を同時に表現するとき，上付き文字を省略して X_{gS} と書くことにする．

現行では，それぞれ観測された $X_{1W}^{(1)}$, $X_{1J}^{(1)}$, $X_{1G}^{(1)}$ から個別に平均点を算出して，それを比較することによって得点調整を行うかどうかの資料となっている．しかし，そこで生じた平均点差は，当然のことながら，科目の難易度の違いから由来するのか，受験者グループの能力の違いから由来するのか判断することができない．つまり，平均点差が，受験者グループの違いを原因としているならば，その差が 50 点であっても得点調整する必要はないだろう．一般的には，科目間に生じた平均点差 D は，以下のように分解することができるだろう．すなわち

$$D = D_S + D_G$$

である．ここで，D_S は科目間の難易度の違いであり，D_G は受験者グループの能力平均の違いである．現行では，D が 20 点差がついたら，得点調整を行うというものであるが，D_S において何点差がついたら得点調整を行うという議論も重要である．

科目選択グループごとに，学力層が異なるかどうかを検討するための基礎資料として，グループ別平均点が参考になる．表 3 は，平成 17 年度の「世界史 B」「日本史 B」「地理 B」選択グループごとの「国語 I・II」「数学 I・A」「数学 II・B」「英語」の 4 科目についての度数・平均・SD である．平成 17 年度は（通年的な傾向でもあるが），「世界史 B」「地理 B」選択者は，各 4 科目の平均点が良いので，「日本史 B」選択者よりも学力層が高いといえるだろう．

このようなグループ別平均点は，科目選択グループごとの学力層の違いを記述している基礎資料であり，実際の得点調整のときにも参考とされている資料である．ただし，学力を代表している科目が英語であるのか，国語であるのか，数学であるのか，あるいはそれらの合計点であるのかを判断するのは難しく，また，国語と数学で必ずしも一貫した結果が得られるとも限らないので，常に読み取りやすい指標ではない．

この問題を解決するためには，表 2 において，それぞれ $X_{1W}^{(1)}$, $X_{1J}^{(1)}$, $X_{1G}^{(1)}$ から算出された平均点を比べるのではなく，可能であるならば $[X_{1W}, X_{2W}, X_{3W}, X_{4W}]$ から算出された「世界史 B」の平均点と，$[X_{1J}, X_{2J}, X_{3J}, X_{4J}]$ から算出された「世界史 B」の平均点と，$[X_{1G}, X_{2G}, X_{3G}, X_{4G}]$ から算出された「世界史 B」の平均点を比較するべきである．そのような受験者集団の能力平均の影響が除去された平均値同士を比較することができれば，残った平均点差は科目間の難易度の差 D_S によるものと言うことができる．

そのような平均（と SD, 相関）を得るには完全情報最尤法（full-information maximum likelihood, FIML）が使える．詳しくは，Arbuckle (1996), Bock, Gibbons & Muraki (1994), Jöreskog & Moustaki (2001), Lee, Poon & Bentler (1990), Muraki & Carlson (1995) などを参照してもらい

表 3 「地理歴史」B 科目選択グループ別「国語・数学・英語」の統計量（平成 17 年度）

選択科目		国 I・II	数 I・A	数 II・B	英語
度数	世界史 B	88,042	58,879	52,171	93,413
	日本史 B	141,408	87,054	75,629	151,597
	地理 B	103,730	101,975	97,574	109,617
平均値	世界史 B	126.1	68.1	50.9	124.6
	日本史 B	120.9	65.6	48.9	115.6
	地理 B	122.8	75.5	56.4	122.8
SD	世界史 B	31.6	23.3	19.0	37.6
	日本史 B	32.2	23.3	18.3	36.5
	地理 B	29.6	20.5	18.1	36.0

たい．FIML は，分布の仮定を必要とする．たとえば多変量正規分布を仮定するならば，各サンプルに対して観測された変数のみを対象とした小さな多変量正規分布を仮定して，全体の尤度を構成する方法である．なお，荘島・大津・田栗 (2006) では，平成 17 年度の現役生を対象に，仮に全ての現役生が主要 16 科目を受験したならば得られるであろう平均，SD，相関を示した．

FIML を用いるには通常，2 つの前提を必要とする．それらは，(1) 多変量正規性，(2) ランダム欠測の仮定である．(1) は，データが多変量正規分布に従うことを要請する．FIML それ自体は，多変量正規性が必要なのではなく，何かの分布を仮定する必要がある．本研究では，多変量正規分布を仮定する必要があるということであり，目的に応じて仮定する分布は変わってよい．多変量正規分布は個々の科目に対しては 1 変量の正規分布に従う．しかしながら，個別科目が正規分布の仮定をすることが難しい場合があることを考えると，多変量正規性を受け入れることは難しいときがあるだろう．ただし，現実的に多変量正規分布に代用できる多変量分布を見つけることは非常に難しく，また仮に存在しても性質がよくわからない分布を使うのは躊躇されるので，多変量正規分布以外の分布を仮定することは事実上困難である．

また，(2) は，欠測メカニズム（Rubin, 1976; Little & Rubin, 1987; 岩崎，2002）がランダム欠測（missing at random, MAR）であると仮定するものである．これは，データが欠測したかどうかの情報を観測データが保持していることを要請するものである．従来，欠測のあるデータを分析しているとき，リストワイズ削除やペアワイズ削除が多用されてきた．これは，欠測メカニズムが完全ランダム欠測（missing completely at random, MCAR）であることを暗に仮定している．しかし，MAR の仮定は，MCAR の仮定に比べると非常に緩やかな制約であり，平均や SD や相関などを推定するに当たり，全てのデータを用いるという当然の要請を担保するという意味合いも含んでいる．荘島・大津・田栗 (2006) はセンター試験データに対して MAR を仮定することについて肯定的であり，「欠測が，情報のある欠測（informative missing, IM）である可能性を疑いつつも，現実にそのときの欠測構造を同定することは難しく，また，多変量データの場合には，変数の数が大きいほど，欠測の情報を他の観測された変数の値が保持している期待が高まるので，MAR の仮定は必ずしも強すぎる仮定とは考えられない．つまり，IM のもとで分析する必要性を頭の片隅に持ちつつも，ソフトウェアの実装状況を考えて MAR を仮定するのが現実的ではなかろうか」と言っている．また，Schafer (1999) は，MAR が成り立つための経験則をいくつか例示しており，その 1 つに “double sampling” がある．Double sampling は，一部の変数は受験者の全てにデータがあればよく，センター試験では，「英語」「国語」「数学 I・A」の科目選択率が高いので，おおむね double sampling の状況が成立していると考えられる．

AMOS 5.0（SPSS, 2003）を用いた．受験者数が 1,000 人を満たない小規模科目の「数学②」教科における「工業数理」「簿記」「情報関係基礎」，および「外国語」教科の「独語」「仏語」「中語」「韓語」は分析から除外した．また，平成 9，10 年度は「旧課程理科①」という科目が存在したが，ここで

表 4 FIML による平均

科目＼年度	H9 (1997)	H10 (1998)	H11 (1999)	H12 (2000)	H13 (2001)	H14 (2002)	H15 (2003)	H16 (2004)	H17 (2005)
国語 I	160.2	150.3	146.2	140.1	126.2	120.1	118.1	131.8	128.8
国語 I・II	137.7	113.7	104.8	110.2	99.8	110.7	99.3	112.1	117.6
世界史 A	55.5	54.2	50.0	48.2	51.9	48.8	48.8	51.7	51.5
世界史 B	65.3	62.6	61.0	61.7	59.0	57.3	54.1	58.8	60.5
日本史 A	53.1	51.3	54.5	49.8	52.5	47.9	47.8	51.4	60.6
日本史 B	55.6	62.4	59.6	62.7	56.2	59.4	64.6	57.1	59.6
地理 A	60.6	62.3	53.5	48.7	63.4	59.1	50.1	62.3	67.9
地理 B	66.0	75.5	60.3	56.0	61.3	63.8	52.8	59.9	68.0
現代社会	56.0	65.1	53.2	43.0	55.0	58.9	57.6	55.3	68.5
倫理	67.3	65.3	56.4	49.7	63.2	60.9	55.5	64.2	61.7
政治・経済	63.2	70.6	56.7	58.3	53.8	52.3	63.3	60.8	63.4
数学 I	49.3	34.4	47.6	71.2	56.0	54.2	47.4	58.5	55.3
数学 I・A	65.2	61.5	48.6	71.7	62.5	61.3	58.8	68.0	67.0
数学 II	47.2	30.7	44.4	41.2	46.4	41.0	39.8	35.9	43.9
数学 II・B	60.6	38.0	57.3	52.1	64.1	54.8	46.4	42.0	48.6
総合理科	69.1	60.5	68.0	55.5	71.3	68.2	61.7	58.1	49.1
物理 I・A	70.8	69.1	70.9	63.8	76.8	75.5	63.9	65.7	70.7
物理 I・B	66.8	59.9	60.4	50.6	65.9	53.7	55.7	57.4	53.7
化学 I・A	55.2	62.6	60.9	67.2	66.2	66.4	59.0	64.4	66.2
化学 I・B	58.4	60.3	57.5	51.3	51.5	51.0	55.6	48.8	59.8
地学 I・A	62.6	56.1	61.2	56.4	66.8	59.4	54.6	60.2	58.7
地学 I・B	63.9	56.8	63.5	62.2	60.1	58.1	53.7	58.8	58.9
生物 I・A	60.9	60.4	59.0	63.1	65.4	72.5	60.5	66.7	58.3
生物 I・B	50.6	62.0	72.5	69.7	66.7	62.0	66.1	61.3	50.6
英語	137.3	127.6	111.3	119.5	110.6	109.6	126.7	130.0	116.0

は分析から除外した．したがって，各年度とも 25 科目による 25 変量正規分布を仮定した．FIML によって推定された結果を表 4 に示す．表 4 には点推定値のみを示しているが，実際には標準誤差も推定されている．ただし，標準誤差は，大標本のためほとんど 0 に近かったためここでは記載しない．参考資料として，分析に利用した各年度の受験者数を示しておく．分析に利用した受験者数であるため，大学入試センターの公式発表である受験者数とは若干異なる．また，表 6 の FIML による平均と実際の平均を平成 9〜17 年度について教科ごとにまとめたものを付録 B に付した．

図 B–1（付録 B）によると，「国語 I」の平均（実線）は，「国語 I」の FIML 平均（破線）に比べて，10 点くらい平均点が低い．なぜなら，「国語 I」の平均は，当該科目を選択した受験生の得点から産出されるいわゆる通常の平均であるが，FIML 平均は，多変量正規性と MAR のもとで，仮に全受験生が「国語 I」を選択したら得られるであろう平均点だからである．つまり，もともと「国語 I」を受験する受験者層は，「国語 I・II」を選択する集団より学力水準が低いことが影響している．また，「英語」の実線と破線がほとんど変わらないのは，「英語」はほとんどの全ての受験生が選択するからである．

そのようにして他の教科も解釈することができる．たとえば，図 B–5 の「数学」では，主要科目である「数学 I・A」および「数学 II・B」の平均点が上がるのは，この科目で比較的学力の高い学生が選択していたためであろう．反対に「数学 I」と「数学 II」は少人数科目であり，もともと学力がそれほど高い受験生が選択しないので，仮に全受験生が選択したとすると，その平均点は上がるため，FIML 平均が高めに推定されたのである．

表 5 分析に用いた各年度の受験者数

科目＼年度	H9 (1997)	H10 (1998)	H11 (1999)	H12 (2000)	H13 (2001)	H14 (2002)	H15 (2003)	H16 (2004)	H17 (2005)
国語 I	58,981	65,338	74,685	81,060	76,142	68,499	54,966	49,452	45,183
国語 I・II	460,039	448,777	426,001	424,423	434,001	457,105	472,248	461,159	450,014
世界史 A	5,277	4,170	2,972	2,811	2,882	2,751	2,331	2,271	2,049
世界史 B	116,741	112,267	102,669	105,272	107,381	109,135	106,583	100,486	93,785
日本史 A	9,021	5,925	4,485	5,275	6,342	6,238	4,242	5,923	5,155
日本史 B	179,571	167,442	153,286	149,309	147,993	154,902	157,610	154,808	152,106
地理 A	7,847	6,210	7,456	7,688	6,985	7,864	7,807	7,861	7,667
地理 B	121,462	121,606	118,547	120,739	128,097	127,271	127,439	119,555	109,827
現代社会	45,957	89,399	158,525	165,339	142,141	164,978	213,101	215,859	198,788
倫理	35,166	50,153	47,741	40,916	34,476	37,110	37,203	48,993	51,453
政治・経済	49,009	52,436	54,554	56,603	59,151	58,136	46,599	65,439	64,269
数学 I	24,845	18,034	17,822	20,493	19,559	18,729	18,433	15,888	13,433
数学 I・A	287,116	374,862	371,758	369,965	381,629	393,382	396,260	382,234	370,253
数学 II	19,293	21,070	15,740	15,497	12,484	13,876	13,474	13,568	11,740
数学 II・B	247,131	322,333	326,955	326,054	336,527	344,301	346,509	339,274	326,762
総合理科	1,276	3,163	4,204	5,212	4,861	8,826	13,040	53,989	78,967
物理 I・A	2,869	2,338	2,449	2,271	2,014	2,205	2,116	2,062	1,746
物理 I・B	158,506	158,702	152,715	151,016	149,738	151,621	148,841	149,678	140,565
化学 I・A	4,769	4,794	5,394	4,784	5,202	5,490	5,793	6,875	6,177
化学 I・B	186,493	186,198	183,928	188,697	190,328	199,102	201,864	219,538	209,900
地学 I・A	3,595	5,447	5,611	5,225	6,656	6,834	7,611	5,101	3,812
地学 I・B	21,537	21,997	19,227	21,585	20,375	24,264	23,656	19,390	18,805
生物 I・A	6,607	6,304	6,667	7,122	7,696	7,569	9,410	10,084	8,709
生物 I・B	155,274	152,900	145,069	151,318	162,551	172,307	176,297	182,744	176,891
英語	548,858	545,182	527,543	529,091	535,336	549,384	552,162	536,213	520,221

得点調整対象科目である「地理歴史」の B 科目（図 B–3）であるが，まず，「世界史 B」「地理 B」は，実線のほうが破線より例年高く，「日本史 B」は実線と破線がほぼ同じ水準であるが，やや破線のほうが高い．つまり，例年，「世界史 B」「地理 B」は比較的学力層が高い集団が選択しているということである．表 3 の結果は，この事実の 1 つの証拠である．すなわち，「世界史 B」「地理 B」選択者は，「日本史 B」選択者よりも学力層が高い傾向がある．実際に得点調整が行われた平成 10 年度について考察すると，表 1 より，「日本史 B」と「地理 B」の平均点差は約 21 点であったが，表 4 よりそれらの FIML 平均点差は，約 13 点であった．これは，例年のように「地理 B」選択者の学力水準は比較的高く，「日本史 B」選択者の学力水準が相対的にやや低いことも影響している．つまり，実際の平均点差は 21 点差であったが，多変量正規性と MAR の仮定の下で，科目間の難易度による差は 13 点ほどであったことがわかる．つまり，21 点のうち 8 点は受験者集団の能力平均に由来していると考えられる．

本小節での主張は，科目選択によって既に学力層にある一定の偏りがあるのだから，平均点が離れても，それは学力層の違いから由来するのか，選択科目の難易度の差から由来するのかが明らかでないということである．しかし，FIML を用いると，多変量正規性と MAR の仮定の下で，仮に全受験生が受けたときの平均点同士を比較することができ，そのとき，科目間の難易度の差を比較することが可能である．

つまり，受験者の能力平均の影響を除去して，純粋に科目間の難易度の差を知りたいときには，通常の平均点を比較するよりも FIML 平均を比較するほうが目的に適っている．その目的においては，

通常の平均点よりもFIML平均のほうが真値に近い値である．なぜなら，通常の平均点は，当該科目を選択した受験者のみから算出されるものだからである．しかし，通常の平均点は，それゆえに値が不変であるが，FIML平均は前提によって値が変わってくる．たとえば，今回は25変量正規分布を仮定したが，A科目などの少数科目をさらに除外して計算することもでき，そのとき仮定される正規分布は25変量よりも少ない多変量正規分布である．そのときの結果と今回の25変量正規分布の元での結果では値が微小に異なってくる．モデルに基づく分析は，このように，仮定するモデルが異なると結果が変わってくる欠点がある．

さらに，FIMLによる結果は，多変量正規分布を仮定するパラメトリックな分析であり，実際に多変量正規性の仮定が成立するのかについては，厳密に言えば疑うことができる．また，欠測メカニズムがMARであるという仮定も成立しているとは考えにくい．そもそも仮定とは，事実とは異なるものであり，成立しているものではない．しかしながら，多重指標における1つの資料となりうる．ただし，MADやRMSDなどと違って一次的な統計資料が得られれば誰でも計算できるものではない．なお，FIMLによって，平均のほかにもSD（分散）と相関（共分散）が推定されている．ここでは掲載しないが，一般的には全受験者が受けていると仮定してもとめたFIMLによるSDが通常のSDよりも大きくなる．また，それに伴って相関係数も大きくなるのが一般的である．

3 シミュレーション

本節では，従来の平均点差に加え，SD・歪度・尖度，さらには，第2節で列挙した十分位点差（DSD），絶対平均偏差（MAD），偏差平方平均平方根（RMSD），FIML平均点差についてそれぞれ基準を設け，「地理歴史」の「世界史B」「日本史B」「地理B」の3科目について，実際の得点調整の要否についてシミュレーションを行う．表6は，8つの指標の絶対値の最大値を示したものである．たとえば，平成9 (1997) 年度，MADの指標が最も大きかったのは「日本史B」と「地理B」の間であり，その値は0.552であった．

続いて，それら8指標による多重指標モニタリングを行う．ただし，多重指標モニタリングを運用するためには，それぞれの指標について基準を決める必要がある．それぞれの指標について基準を決める際，どのように定めるかについては1つ1つの指標についてそれぞれ吟味する必要があり，それぞれの基準の決定は簡単ではない．

本節では，以下のように，それぞれの指標について3水準を設けた．それらは，

- 平均点差：10, 15, 20
- SD差：4.0, 5.0, 6.0
- 歪度差：0.2, 0.4, 0.6
- 尖度差：0.50, 0.75, 1.00
- 十分位点差：5, 10, 15
- MAD：0.50, 0.75, 1.00
- RMSD：1.0, 1.5, 2.0
- FIML平均点差：5, 10, 15

である．いずれも数値が小さいほど，分布が類似しているための厳しい基準となっている．上記のような設定の下で，実際に平成9 (1997)～17 (2005) 年度の「世界史B」「日本史B」「地理B」において，得点調整の要否を検討する．結果を表7に示す．

表7において，0は指標が基準を満たすこそを表し，1は基準を超えたことを表している．たとえば，平成9年度において，指標の基準を (平均点差，SD差，歪度差，尖度差，十分位点差，MAD，RMSD，FIML平均点差)=(15, 5.0, 0.4, 0.75, 10, 0.75, 1.5, 10) と設定するならば，判定結果は (0, 0, 0, 0, 1, 0, 1, 1) となる．つまり，十分位点差，RMSD，FIML平均点差の3つの基準に抵触したことを示している．ただし，3つの指標に抵触したとしても，4つ以上の指標に抵触した場合に得点調整を行うというルールを設定すれば，得点調整を行わないという選択肢もありうる．

表7より，平均点差=20という基準を満たしていても，他の基準に抵触している年度が存在することが分かる．比較的，平成11，15，16年度は抵触している基準の数が他の年度に比べて相対的に少ない．これは，付録Aの図A–3，A–7，A–8から見て分かるとおり，他の年度に比べて比較的分布の形状と位置が類似しているからである．

第2節で提案した指標の性質を見てみる．表7より，十分位点差が5点以内におさまることはなかった．十分位点差は，分位数が9つある．そのとき，比較する科目数が3つのとき27（=9分位

表 6 8 つの指標の最大値の結果

	H9 (1997)	H10 (1998)	H11 (1999)	H12 (2000)	H13 (2001)	H14 (2002)	H15 (2003)	H16 (2004)	H17 (2005)
平均点差	12.50 世日	20.90 日地	5.15 日地	6.41 世地	8.29 日地	7.62 日地	8.93 日地	5.59 日地	10.94 日地
SD 差	4.29 世地	4.25 世地	5.40 世地	6.43 世地	6.34 世地	4.11 世地	3.95 世地	3.57 世地	5.34 世地
歪度差	0.236 日地	0.619 日地	0.194 日地	0.219 世地	0.434 日地	0.457 日地	0.212 日地	0.160 日地	0.445 世地
尖度差	0.466 日地	1.151 世地	0.436 世地	0.594 世地	0.737 世地	0.571 世地	0.356 世地	0.454 世地	0.896 世地
十分位点差	15 世日 日地	25 世地 日地	8 世日 世地	15 世地	15 日地	13 世地	13 日地	9 日地	14 世地
MAD	0.552 日地	1.010 日地	0.460 世地	0.579 世地	0.561 日地	0.396 日地	0.486 日地	0.309 世地	0.540 日地
RMSD	1.520 日地	2.875 日地	1.302 世地	1.682 世地	1.597 日地	1.156 日地	1.401 日地	0.825 世地	1.556 日地
FIML 平均点差	10.39 日地	13.09 日地	1.42 世日	6.72 日地	5.12 日地	6.45 世地	11.81 日地	2.73 日地	8.34 日地

表 7 多重指標モニタリングによる判定結果
（基準を超えたら 1，基準内であれば 0）

指標	基準	H9 (1997)	H10 (1998)	H11 (1999)	H12 (2000)	H13 (2001)	H14 (2002)	H15 (2003)	H16 (2004)	H17 (2005)
平均点差	10	1	1	0	0	0	0	0	0	1
	15	0	1	0	0	0	0	0	0	0
	20	0	1	0	0	0	0	0	0	0
SD 差	4.0	1	1	1	1	1	1	0	0	1
	5.0	0	0	1	1	1	0	0	0	1
	6.0	0	0	0	1	1	0	0	0	0
歪度差	0.2	1	1	0	1	1	1	1	0	1
	0.4	0	1	0	0	1	1	0	0	1
	0.6	0	1	0	0	0	0	0	0	0
尖度差	0.50	0	1	0	1	1	1	0	0	1
	0.75	0	1	0	0	0	0	0	0	1
	1.00	0	1	0	0	0	0	0	0	0
十分位点差	5	1	1	1	1	1	1	1	1	1
	10	1	1	0	1	1	1	1	0	1
	15	1	1	0	1	1	0	0	0	0
MAD	0.50	1	1	0	1	1	0	0	0	1
	0.75	0	1	0	0	0	0	0	0	0
	1.00	0	1	0	0	0	0	0	0	0
RMSD	1.0	1	1	1	1	1	1	1	0	1
	1.5	1	1	0	1	1	0	0	0	1
	2.0	0	1	0	0	0	0	0	0	0
FIML 平均点差	5	1	1	0	1	1	1	1	0	1
	10	1	1	0	0	0	0	1	0	0
	15	0	0	0	0	0	0	0	0	0

数×3ペア），科目数が4つのときは54（=9分位数×6ペア）も参照する必要がある．それらの全てが5点以内に収まることは難しい．したがって，十分位点差が，もっとも厳しい指標となっている．しかしながら，十分位点差は，グループを十分割する点であるので，ある得点層において不利益が生じていないかどうかを直接的に判断することができる指標である．したがって，十分位点差の基準を甘くすることは，特定の得点層における不利益を見過ごすことになりかねない．十分な議論を行って決める必要があるだろう．

基準の設定の仕方に依存するが，シミュレーションの設定の中では，MADは，平均点差の次に鈍感な指標であった．また，MAD=0.5だと検出しやすいが，0.75であると検出しにくいので，その間がよいかもしれない．同様な理由から，RMSDは，1.5から2.0の間くらいがよいかもしれない．引き続いて議論が必要である．

FIML平均点差の基準を10に設定すると，過去9年間では3回抵触することになる．FIML平均は，MARと多変量正規性のもとで，仮に全受験生が，当該科目を受験したら得られるであろう平均点であるので，そのモデルの下では，能力差を除去した平均点の推定値である．つまり，推定された平均値は，そのモデルの下で問題の難易度の差によるものである．したがって，100点満点のテストで難易度だけで10点以上差がつくことは，特定の受験生の不利益になりかねないので，やはりFIML平均点差の基準を10点以上にすることは難しいだろう．今後の議論が必要である．

4 おわりに

本研究では，多重指標モニタリングという得点調整手続きの枠組みを提案した．多重指標モニタリングを運用するには，①指標の選択，②それぞれの指標の基準の設定，③いくつの指標に抵触したら得点調整をするか，という大きく3つのパラメタを設定する必要がある．①については，従来の平均点差以外に，SD差，歪度差，尖度差，十分位点差，絶対平均偏差（MAD），偏差平方平均平方根（RMSD），完全情報最尤法（FIML）による平均点差が有用であると提案した．しかし，今回は取り上げていない指標，あるいはより良い指標を作ることも今後の課題となろう．②については，各基準について暫定的に3水準を設定し，シミュレーションを行ったが，引き続き十分な検討を行う必要がある．基準は，どこかに設定しなくてはいけない．ノンパラメトリック検定を用いることができたとしても（センター試験の事例では，標本サイズが大きいので使えないが），有意水準はどこかに定めなくてはならない．合意を得るまで議論をする必要があるだろう．③については1つでも抵触すれば，得点調整を行うという考え方から，全ての指標に抵触したら得点調整を行うということまで考えられる．無論，③は，②と密接に関係しているので，独立に設定することができない．

得点調整は，受験生の利害に絡む重要な問題であり，ひいては，わが国の人材配置にかかわる問題である．本研究での結論を急がず，得点調整は毎年やるべきである，という根本的な意見も含めて，十分に時間をとって議論をすべきである．

謝　辞

査読者の先生方には，大変つたない初稿を丁寧に読んでくださってコメントをいただきました．おかげで改稿では，ほとんど別の論文かのように見違えました．この場をお借りして深く感謝申し上げます．

参考文献

Arbuckle, J.L. (1996). Full information estimation in the presence of incomplete data. In G.A. Marcoulides & R.E. Schumacker (Eds.) *Advanced Structural Equation Modeling: Issues and Techniques*. Lawrence Erlbaum Associates, Inc. (pp.243–277).

Bock, R.D., Gibbons, R., & Muraki, E. (1994). Full-information item factor analysis. *Applied Psychological Measurement*, **12**, 261–280.

大学入試センター (2006). 平成19年度センター試験実施要綱〈http://www.dnc.ac.jp/center_exam/19exam_index.html〉（2006年10月29日）

福嶋裕・高橋美貴・樽磨和幸・佐藤仁 (2005). 試験の得点調整に関する提案 大阪商業大学論集，**138**, 1–11.

岩崎学 (2002). 不完全データの統計解析 エコノミスト社.

Jöreskog, K.G. & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, **36**, 347–387.

Lee, S.-Y., Poon, W.-Y., & Bentler, P. (1990). Full maximum likelihod analysis of structural equation models with polytomous variables. *Statistics & Probability Letters*, **9**, 91–97.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

前川眞一 (2003). 分位点差縮小法を行うための SAS 擬似プロシジャ PROC1 RPM の使い方 Version 3.0 大学入試センター研究開発部リサーチノート, RN-02-15.

Muraki, E. & Carlson, J.E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, **19**, 73–90.

Rosenbaum, P.R. & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of American Statistical Association*, **79**, 516–524.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

荘島宏二郎・大津起夫・田栗正章 (2006). 自己組織化マップと完全情報最尤法から見た平成 17 年度センター試験の教科学力構造日本テスト学会誌, **2**, 75–89.

SPSS, Inc. (2003). *AMOS 5.0*. (http://www.spss.co.jp/).

SPSS, Inc. (2005). *SPSS 14.0*. (http://www.spss.co.jp/).

柳川堯 (1982). ノンパラメトリック法 培風館.

付録 A

付録では，平成 9 (1997)～17 (2005) の「地理歴史」教科における「世界史 B」「日本史 B」「地理 B」の得点分布（相対度数分布）を示す．

図 **A–1** 平成 9 (1997) 年度

図 **A–2** 平成 10 (1998) 年度

図 **A–3** 平成 11 (1999) 年度

図 **A–4** 平成 12 (2000) 年度

図 **A–5** 平成 13 (2001) 年度

図 **A–6** 平成 14 (2002) 年度

図 **A–7** 平成 15 (2003) 年度

図 **A–8** 平成 16 (2004) 年度

図 **A–9** 平成 17 (2005) 年度

付録 B

付録 B では，表 5 に基づく FIML による平均と，通常の平均を同時にプロットしたものを教科のカテゴリごとに図示したものを掲載する．実線は，いわゆる平均値（図中，M）を示し，破線は，FIML によって得られた平均（図中，F）を示している．

図 **B–1** 「国語」「外国語」の平均と FIML 平均

図 **B–2** 「地理歴史」A 科目の平均と FIML 平均

図 **B–3** 「地理歴史」B 科目の平均と FIML 平均

図 **B–4** 「公民」の平均と FIML 平均

図 **B–5** 「数学」の平均と FIML 平均

図 **B–6** 「理科」における「総合理科」と A 科目の平均と FIML 平均

図 **B–7** 「理科」B 科目の平均と FIML 平均

A score adjustment procedure using multiple indices monitoring

SHOJIMA Kojiro*
ISHIZUKA Tomoichi*
HASHIMOTO Takamitsu*
OTSU Tatsuo*
MAYEKAWA Shin-ichi**

Abstract

By analyzing test-score data from 1997–2005 for students who took the National Center's "Geography and History" test, it was learnt that the mean difference data was insufficient to discover the difference in distributions. We propose a new score adjustment procedure, "multiple indices monitoring," that can be used to detect the difference of distributions by analyzing two or more indices such as the SD differences, the skewness differences, the kurtosis differences, the decile score differences, the mean absolute differences, the root mean square differences, and the mean differences using the full-information maximum likelihood method.

Key words: score adjustment, geography and history, missing at random (MAR), full-information maximum likelihood method.

* Research Division, The National Center for University Entrance Exainations
** Graduate School of Decision Science and Technology, Tokyo Institute of Technology