【原著】 大学入試研究ジャーナル第 32 号，122-129，2022

傾向スコアを用いた令和３年度共通テスト公民科目間差の試算

――国語・英語リーディング・英語リスニングを共変量に用いた場合――

荘島 宏二郎，橋本 貴充，宮澤 芳光，石岡 恒憲，前川 眞一（大学入試センター）

令和 3 (2021) 年 1 月に実施された第 1 回大学入学共通テストでは，公民 3 科目（現代社会，倫理，政治・経済）において得点調整が実施された。3 科目の最大平均点差が 20 点以上つき，かつ，その差が試験の難易差に基づくと判断されたためである。しかし，公民 3 科目の受験者集団はそれぞれ別集団であり，3 科目の平均差には受験者集団の学力差も反映している。本研究では，傾向スコアを用いて集団間の学力差を事後的に調整した上での科目間平均差（分布差）について報告した。分析の結果，18.7 点あった科目間平均差が 14.8 点まで縮小した。したがって，令和 3 年度公民科目の難易差は大きく不揃いでなかったと言える。しかし，分析のときに採用した各種の仮定や条件を鑑みたとき，難易差が 15 点であると断定することはできず，現実に 20 点以上の平均差がついている状況下で，得点調整を行ったという判断は適切であったと思われる。

キーワード：大学入学共通テスト，公民，得点調整，傾向スコア

1 はじめに

令和 3 (2021) 年 1 月，第 1 回の大学入学共通テスト（以下，共通テスト）が実施された。当年度の共通テストは，新型コロナの影響により，第 1 日程（1 月 16・17 日），第 2 日程（1 月 30・31 日），および特例追試験（2 月 13・14 日）という，通常の全 2 回 4 日間日程とは異なり全 3 回 6 日間日程が用意された。第 1 日程と第 2 日程は本試験の扱いであった。

多くの受験者は第 1 日程に志願し（約 53 万人），そのうち，何らかの科目を受験した人数は 482,546 人であった（大学入試センター, 2021a）。

共通テストでは 6 教科 31 科目が用意されているが，以下の 3 つの科目群

地理歴史：世界史 B，日本史 B，地理 B
公民　　：現代社会，倫理，政治・経済
理科②　：物理，化学，生物，地学

において，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, 2020）とされている。ただし，受験者数が 1 万人に満たない科目については，得点調整対象科目から外れるという付則がある。

得点調整は，科目選択が本人希望によるものとはいえ，あまりにも科目間難易度に差が生じた場合，難易度の高かった科目を選択した受験者の出願行動に対して不利益が生じかねないため，それを是正するために行うものである。2021 年度共通テストでは，第 1 日程の公民 3 科目間において、平均点差が 20 点以上つき，その差が難易差に基づくものと判断されたため，得点調整が実施された（大学入試センター, 2021b）。また，第 1 日程の理科②においても，物理・化学・生物の 3 科目において平均点差が 20 点以上開き，かつ，その差が難易差に基づくものと判断されたため，得点調整が実施された（大学入試センター, 2021c）。ただし，地学は受験者数が 1 万人に満たないため，得点調整対象から除外された。

上述したように，得点調整は，原則として科目間平均点差が 20 点以上つき，これが試験問題の難易差に基づくと判断された場合に実施される。実際は，大学入試センター試験時代も含めて，過去に平均点差が 20 点以上離れたときは，試験問題の難易差に基づくと判断され，常に得点調整が発動された経緯がある。しかし，本来は，平均点差が 20 点以上ついたとしても，それが試験問題の難易差に基づかないと判断されたならば，得点調整は行わないということもありえる。

表 1 公民の受験者数・平均点（中間集計）

	受験者数	平均点
現代社会	21,217	54.34
倫理	6,344	71.76
政治・経済	15,779	51.32

一口に，試験の難易差といっても，考え方は一通りではない。荘島ら（2007）のように，科目間の中央値差が 20 点ついたとき難易差があるとする判断もありえる。また，この問題を考えるとき，問題を複雑に

しているのが，受験者集団が同一でないという事実である。例えば，当年度の第１日程では，公民３科目の中間集計[1]は，表１の通りであった（大学入試センター, 2021d）。

表１より，倫理の平均点が最も高く，現代社会（以下，現社）の平均点が続き，政治・経済（以下，政経）の平均点が最も低かったことが分かる。しかし，３科目の受験者数から分かる通り，部分的に受験者は重複するものの，それぞれの科目を選択した受験者は異なるため，倫理の試験が簡単だったため平均点が高かったのか，倫理受験者集団の学力が高かったから平均点が高かったのか簡単には判断できない。

したがって，何らかの方法を用いて，公民３科目の受験者集団を均質化し，そのもとで平均点差などを比較することは重要である。そのような試みは，非線形因子分析を用いた大津（2011）やランダムフォレストを用いた石岡（2011）にも見られる。しかし，それらの方法は一般的に高度な手法であり，多くのソフトウェアで標準的に搭載されている手法ではない。本研究では，簡単に用いることができる傾向スコア（Rosenbaum & Rubin, 1983; 吉村・荘島, 2004; 荘島ら, 2006）を用いて，科目間の難易差を見積もった分析事例を報告する[2]。

2 傾向スコアによる共変量調整

2.1 準実験

傾向スコア（propensity score）は，医療・薬効評価などの分野でしばしば用いる準実験(quasi-experiment)というデータ計画で用いられる共変量調整の１つである。たとえば，ある薬の臨床効果を検証したいとき，様々な理由で集団をランダムに２群に分けられない場合がある（表2）。

表２ 準実験計画

群	偽薬	投薬	共変量
統制群	○	×	○
介入群	×	○	○
効果の平均	$\bar{x}_0$	$\bar{x}_1$	

○観測，×欠測

そのとき，実験群に薬を投与し，予後の経過が良かった（$\bar{x}_1 > \bar{x}_0$）としても，薬効があったと結論付けることはできない。もともと実験群に割り当てられた集団の健康度が高かったからかもしれないからである。この状態は，先述したような，倫理が簡単だったのか，倫理受験者集団の学力が高かったのか判別つかないという状態と構造が類似していることが分かるだろう。

そのとき，実験群・統制群ともに観測されている共変量を用いて，事後的に（統計的に）両群を均質に近づける方法が共変量調整であり，その方法の１つが傾向スコアによる共変量調整である。共変量(covariate)とは，例えば，年齢・性別・血糖値など，薬効に関わるであろう様々な変数が候補であり，実験群・統制群共に観測されていることが肝要である。したがって，例えば年齢・性別・血糖値を共変量と指定したならば，それら共変量に関して均質な（ランダムな）2 群を事後的に作り出すことを意図している。

本研究では，公民３科目の受験者集団がランダム割り当てになっていないために，3 科目の平均点にはテストの難易のほかに受験者集団の学力差が混入している状態である。3 科目の受験者集団がランダムに割り当てられていてこそ，3 科目の平均点差を純粋に比較することができる。そのため，3 科目の受験者集団を事後的に均質化することが本研究の目的である。

2.2 共変量

公民３科目の受験者集団の学力を事後的に均質化する上で，どの変数を共変量に用いるかは重要な決定となる。表３は，本分析計画である。受験者は自由に他科目と組み合わせ選択できるため，受験者の選択行動は様々である。例えば，群４は，地歴公民のうち，現社と倫理を選択回答した受験者集団である。

表３　本分析計画

群	現社	倫理	政経	共変量
群１	○	×	×	○
群２	×	○	×	○
群３	×	×	○	○
群４	○	○	×	○
群５	○	×	○	○
群６	×	○	○	○
平均点	$\bar{x}_C$	$\bar{x}_E$	$\bar{x}_P$	

○観測，×欠測

現社の平均点$\bar{x}_C$（添え字Cはcontemporary societyより）は群1, 4, 5の受験者の集計結果であり，倫理の平均点$\bar{x}_E$（Eはethicsより）は群2, 4, 6の受験者の集計結果であり，また，政経の平均点$\bar{x}_P$（Pはpolitics and economicsより）は，群3, 5, 6の受験者の集計結果である。したがって，群間で大きな学力差があるがゆえに引き起こされた平均点差の可能性が排除できない。

続いて，公民３科目の学力に大きく関与しているであろう科目を共変量に指定すべきであるので，国語，英語リーディング（以下，英語 R），そして英語リスニング（以下，英語 L）が挙げられる。また，共変量の要件として，全受験者に観測されている必要がある。国語・英語R・英語Lは第１日程受験者の大多数が受験しているため（順に 457,305 人，476,174 人，474,484 人），公民３科目の共変量として適切である。

数学Ⅰ・数学Aや数学Ⅱ・数学Bもまた，基礎的学力を構成する重要な科目であるが，公民受験者は必ずしも数学を選択受験しないので，本分析では共変量として不適とした[3)]。

2.3 分析データ

本分析で用いる公民３科目の分布を図１に示す。分布の太さは，標本の大きさを示している。本分析で用いるデータの標本サイズはn=83,335 であった。全受験者数の約 1/5 にあたり，中間集計データから約 1/3）を切り出したデータであり，全体の傾向は変わらない。このデータでは，共変量の全科目を選択していない受験者は除外されている。共変量はすべて観測していることが傾向スコアを推定する上での原則だからである。このデータにおいて，現社・倫理・政経の標本サイズ（平均点）は，順に 7,193 人（56.0），2,534 人（71.3），5,963 人（52.6）であった。

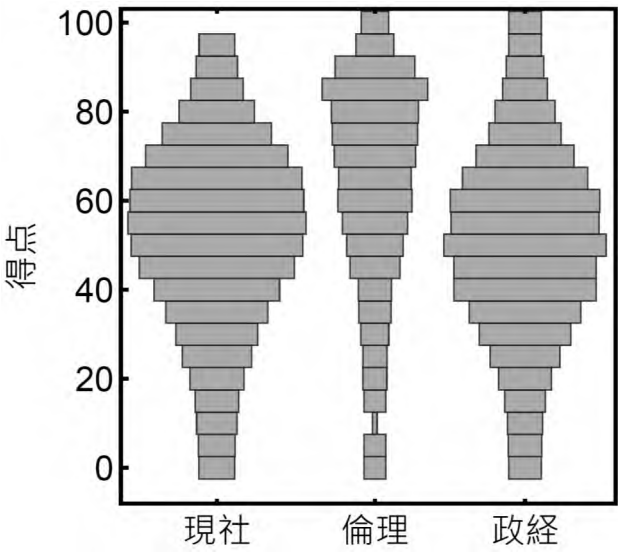

図１ 分析に用いた公民３科目の得点分布

図１から明らかなように，倫理の平均が最も高く，政経の平均が最も低い。このデータでの平均点差は 18.7 点であった。しかし，この結果から倫理のテストが最も易しかったと結論することはできない。なぜなら，これまで倫理選択者は比較的学力が高い集団であることが報告されており（大津, 2011; 石岡, 2011; 橋本, 2021），今回もその可能性が高いと考えられるからである。

2.4 傾向スコア推定と重みづけ

続いて，共変量を利用して，各受験者の各公民科目に対する「選びやすさ」を計算する。この「選びやすさ」こそ傾向スコア(propensity score)と呼ばれ，0 から１の確率（あるいはメンバーシップ）で得られる。言い換えれば，国語，英語R，英語Lの得点から推定される各公民科目を受験する傾向であると解釈できる。この傾向スコアを計算することで，例えば，現社を受けた者もそうでない者も，現社に対する「選びやすさ」を推定できる。

この傾向スコアを推定するときには，さまざまな統計モデルを用いることができるが，本分析では，ロジスティック回帰分析を用いた。例えば，現社の傾向スコアを推定するには，まず，現社を選択した受験者を 1，選択しなかった受験者を 0 とした 2 値変数を従属変数とし，国語・英語R・英語Lを独立変数としたロジスティック回帰分析を行い，ロジスティック回帰係数$\beta_0, \beta_J, \beta_R, \beta_L$を得た（表 4）。

表 4 ロジスティック回帰分析結果

	選択科目		
	現社	倫理	政経
β_0	−1.763**	−4.022**	−1.252**
β_J	0.004**	0.016**	−0.005**
β_R	−0.011**	−0.018**	−0.007**
β_L	−0.007**	−0.004	−0.005**
−2×対数尤度	48511.2	22280.5	42128.5
R^2	0.013	0.021	0.024

** p<.001

倫理選択の 2 値変数を従属変数としたときの英語Lの回帰係数のみ有意でなかったが，その他の係数は全て有意であった。概して，回帰係数は負となっている。これは，公民科目は地歴科目と同時選択可能であり，公民選択者は，比較的学力が低い集団が選択する傾向があるという構造が原因である。ただし，現社と倫理は国語の得点の高い受験者が選択してくる傾向がある。

また，説明率（R^2）は全体的に大きくなかった。しかし，著者らは，共変量の得点パタンによって科目

選択が完全に決まるとまで期待しておらず，共変量が科目選択に与える影響は限定的であろうことは想定していた。しかし，たとえ影響が限定的であろうとも，共変量が科目に及ぼす影響は調整したほうがよいと考え分析を続ける。

そして，受験者iの現社の傾向スコアe_{Ci}を以下の式で推定する。すなわち，

$$e_{Ci} = \frac{1}{1+\exp\{-(\beta_0 + \beta_J x_{Ji} + \beta_R x_{Ri} + \beta_L x_{Li})\}}$$

である。ここで，x_{Ji}, x_{Ri}, x_{Li}は，それぞれ，受験者iの国語・英語R・英語Lの得点である。ロジスティック回帰分析は，多くのソフトウェアによって標準的に装備されている統計手法である。本分析では，SPSS version 25 を用いた。

続いて，この傾向スコアを用いて，現社選択者である受験者iの現社得点x_{Ci}の得点区間tに対する寄与を傾向スコアを用いて以下のように重みづける。

$$f_{Ci}^{(t)} = \begin{cases} e_{Ci}^{-1}, & \text{if } x_{Ci} = t \\ 0, & \text{otherwise} \end{cases} \quad (t = 0, \cdots, 100)$$

このような手法を inverse probability weighting (IPW; 星野, 2009; 菱山・岡田, 2019; 斎藤 2020)という。これは，x_{Ci}が得点区間tに分類されるとき 1，そうでなければ 0 であるという 2 値得点を逆確率で重みづけていることを表す。例えば，傾向スコアが 0.5 であった受験者は得点tにおける頻度の寄与を$2(=1/0.5)$人分とカウントする。傾向スコアが高い受験者の重みはほとんど変えず，傾向スコアが低い受験者ほど大きく重みづけることによって，受験しなかった者を含めた全体の分布を再現しようと試みている[4]。

傾向スコアの値は，国語，英語R，英語Lの得点パタンから推定される現社を受験する傾向であると解釈できた。傾向スコアの値が小さい場合には，その傾向スコアをもつ受験者の国語，英語R，英語Lの得点パタンで現社を選択した受験者が少なかったことを意味する。傾向スコアの逆数による重みづけの意味は，国語，英語R，英語Lの得点パタンが同じであれば，現社を選択した受験者の現社の得点と現社を選択しなかった受験者が仮に現社を受験した場合の得点が同程度であったと仮定し，傾向スコアが小さい受験者の国語，英語R，英語Lの得点パタンに対して，相応する現社の得点をとった受験者数を仮想的に増やしていると解釈できる。

図２に推定された各科目受験者の傾向スコアの分布を示す。全体的に３つの分布は０に近いところで分布しており，傾向スコアが 0.20 以上である受験者がいなかったことを表している。これは，公民科目がそもそも受験者に選択されにくい構造が反映されたものである。

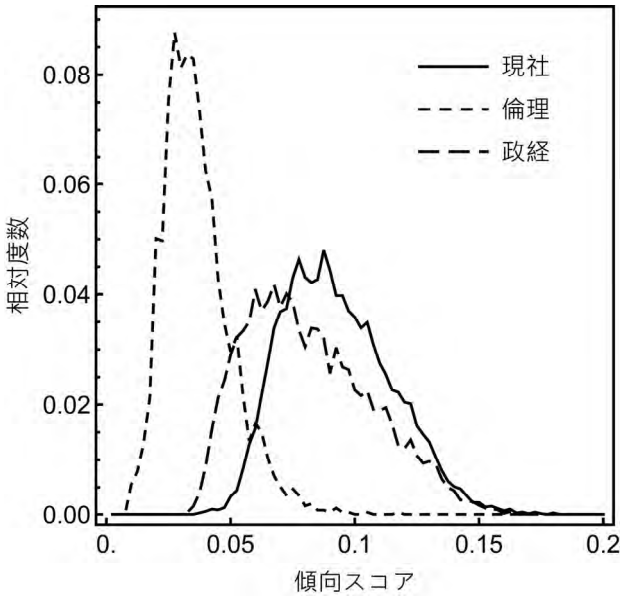

図２ 傾向スコアの分布

傾向スコアの分布の重なりがあまりに小さいとき，事後的に無作為割り当てがなされた matched sample を作成することが難しくなる。しかし，本研究では，matched sample を作成するのではなく，現社選択者の現社の分布から全受験者の現社の分布，倫理選択者の倫理の分布から全受験者の倫理の分布，政経選択者の整形の分布から全受験者の政経の分布を再現することを目的としているため，傾向スコアの分布の重なりは重要でない。とはいえ，あまりにも極端な違いがあると，引き続く分析の推定が不安定になる（Li, Thomas, & Li, 2019）可能性がある。しかし，図２より，傾向スコアが 0.04 から 0.1 の範囲における現社受験者・倫理受験者・政経受験者の割合は 66.7%, 30.5%, 76.1%いるため，３つの分布が乖離しているとは言えないと判断した。

続いて，共変量調整後の度数分布における得点tの期待相対度数は

$$f_C^{(t)} = \frac{\sum_{i=1}^{n} f_{Ci}^{(t)}}{n} \quad (t = 0, \cdots, 100)$$

として推定することができる。したがって，累積相対度数分布において得点τ以下の区間を

$$R_C^{(\tau)} = \sum_{t=0}^{\tau} \frac{f_C^{(t)}}{\sum_{t'=0}^{100} f_C^{(t')}} \quad (\tau = 0, \cdots, 100)$$

として求めることができる。数理的な詳細は，吉村・

荘島（2004）や荘島ら（2006）を参照されたい。

同様の手続きにより，「仮に全受験者が倫理を受けたら」「仮に全受験者が政経を受けたら」，それぞれどのような倫理と政経の分布になるかを試算することができる。そして，これら「仮に全受験者が現社／倫理／政経を受けたときの 3 つの分布」を比較することによって，無作為配置されていない群間の差を事後的に均質化する。

2.5 共変量調整後の分布

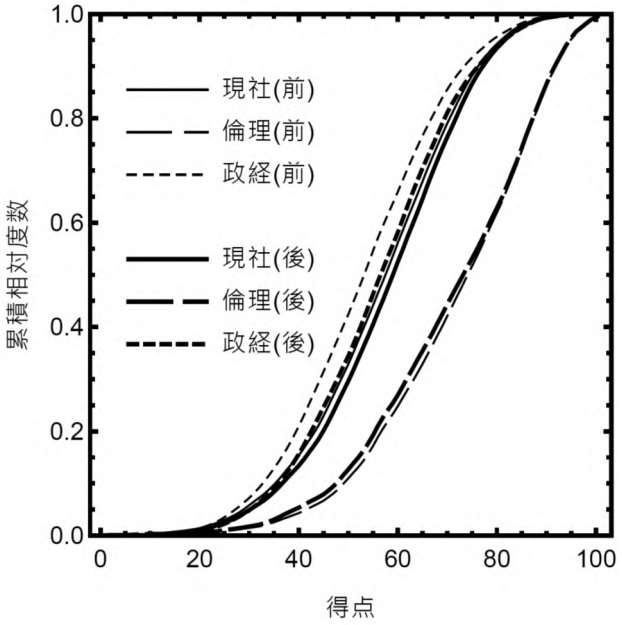

図 3 共変量調整前後の公民 3 科目の累積相対分布

以上のような手続きにより，傾向スコアを用いた共変量調整を行った。共変量調整前後の公民 3 科目の累積相対度数分布を図 3 に示す。3 本の細線が共変量調整前の分布であり，3 本の太線が共変量調整後の分布を示している。一番右に位置している 2 本の分布は，倫理の共変量調整前後の分布である。共変量調整後の分布の方は，主に 50%以下のところで低めに補正されていることが確認できる。これは，倫理の実受験者が，（共変量について，すなわち国語・英語 R・英語 L に関して）学力が高い集団であることを示唆している。「仮に全受験者（分析データの 83,335 人）が倫理を受験すると」，実受験者よりも学力の低い集団が混入してくるので，共変量調整後の分布は低めに補正されたと推察できる。この結果は，大津（2011），石岡（2011），橋本（2021）とも一致した傾向を示している。

また，図中，一番左の 2 本の破線は，政経の分布を示している。細線が共変量調整前，太線が共変量調整後の分布である。共変量調整後の分布の方が高めに補正されていることが確認できる。これは，倫理の実受験者とは逆に，政経の実受験者が，（共変量に関して）学力の低い集団であったことを示唆している。その結果，「仮に全受験者（分析データの 83,335 人）が政経テストを受験すると」，実受験者よりも学力の高い集団が混入してくるので，共変量調整後の分布は高めに補正されたと思われる。

前述したように，共変量調整前，公民 3 科目の最大平均差は 18.7 点であった。しかし，共変量調整後，公民 3 科目の最大平均差は 14.8 点まで縮小した。

2.6 共変量のバランス

本節では，傾向スコアを用いて共変量調整をしたことによる適切さに関する議論の補足を行う。IPW で重みづけしたことにより，例えば，現社選択者の共変量の分布は調整後，全受験者の共変量の分布に理論上一致するはずである。一致度が低いとき，現社選択者から復元する分布はバイアスが残っていると考えられる。バランスの基準としてよく用いられる標準化平均値差（standardized mean difference, SMD）と，分布の差を計量する指標として L^2 距離，Kullback-Leibler（KL）距離について，調整前後でそれらの指標がどのように改善したかについて，表 5 にまとめた。

表 5 調整前後の共変量の分布の一致度

受験者	共変量	調整前			調整後		
		SMD	L^2	KL	SMD	L^2	KL
現社	国語	.0781	.0007	.0607	.0096	.0001	.0097
	英 R	.2254	.0012	.0443	.0134	.0001	.0046
	英 L	.2102	.0007	.0364	.0157	.0001	.0050
倫理	国語	.2571	.0010	.0963	.0040	.0001	.0049
	英 R	.0717	.0006	.0292	.0073	.0001	.0030
	英 L	.0410	.0004	.0188	.0079	.0001	.0030
政経	国語	.3283	.0008	.0772	.0636	.0003	.0285
	英 R	.3415	.0019	.0772	.0114	.0001	.0045
	英 L	.3155	.0012	.0605	.0162	.0001	.0061

例えば，現社選択者の英語 R の分布と，全受験者の英語 R の分布の SMD は，調整前で 0.2254 であったが，調整後は 0.0134 までに改善している。一般的に，SMD は 0.1 を下回ると 2 群の分布がバランスさ

れていると評価される(Austin, 2011)。また，全ての科目選択集団の全ての共変量において，調整後のSMD が 0.1 を下回った。同様に，L^2距離と KL 距離も調整後大きく改善していることが確認できる。このことは，国語・英語Ｒ・英語Ｌを共変量に指定し，これら共変量を用いて傾向スコアを推定し分布を修正することの妥当性の１つの根拠を示している。

2.7 考察

共変量調整前の平均点差（18.7 点）には，テストの難易度に加えて，受験者集団の学力差が混入している。傾向スコアを用いた共変量調整の結果，この差が 14.8 点まで縮小した。得点調整対象科目以外の情報を用いることで，科目間の難易差をより明確に切り出すことができる。

しかし，この 14.8 点差を大きいと判断するか，小さいと判断するかは，分析者・研究者・受験生など立場によって様々であろう。しかし，テスト作成を行った経験があるならば，どんなに均質な平行テストを作ろうと心がけても 100 点満点で±15 点差くらいは珍しくないということは経験的に分かるだろう。

共通テストでは，各テストにつき大学の教員が 20 名ほどで問題作成を行っており，2 年かけて慎重に作成している。どの科目も同じ手間・時間・人数をかけて作成している。それでも同一教科内の異なる科目間で 15 点くらいの差は時々起こるということは，過去のセンター試験の平均点差を眺めれば一目瞭然であろう。その意味では，令和３年度の公民３科目の難易差は，大きく不揃いであったとは言えない。

今回は，全データの 20%ほどの中間集計データでの結果であるため，全データを用いても，平均点差は 20 点以内に収まり 15 点差近くまで縮小すると思われる。したがって，令和３年度の公民では得点調整が行われたが，その 20 点差は「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」という判断基準において，20 点の差すべてが難易差に起因するものではないため，解釈によっては得点調整をしないという判断もありえたかもしれない。しかし，学力差をある条件下で除去してもなお約 15 点差が残り，かつ，現実に平均点差が 20 点以上ついている状況で，受験生の立場（特に現社と政経の選択者）から考えれば，得点調整を行ったという判断は合理的である。しかし，この事例１つとっても得点調整の是非や当否を決めるのは非常に難しい判断を要するということを付言したい。

3 本研究の仮定と限界

最後に，本分析で用いた制約と仮定について整理しておく。これは，とりもなおさず，本研究の限界を示唆するものである。それらは，以下の(1)～(5)である。

(1) 中間集計データ
(2) 共変量は国語・英語Ｒ・英語Ｌ
(3) 共変量の完全解答者
(4) Weak unconfoundedness の仮定
(5) ロジスティック回帰分析モデル

まずは，(1)より，全受験者のデータではなく，約８万人の中間集計データを用いた。全受験者のデータを用いた結果とは異なる。

また，(2)より，共変量として国語・英語Ｒ・英語Ｌの３科目を指定した。先述したが，科目選択による学力差を事後的に均質化するためにこの３科目でなくてはいけないということではない。数学Ⅰ・数学Ａや数学 Ⅱ・数学 B などの主要科目を共変量に追加することも可能である。しかし，表３に示した通り，共変量は，公民科目選択者が普遍的に選択しているような科目である必要があるため，必然と大受験者科目を指定することになる。しかし，いずれにせよ，本分析が示す結果は，共変量に指定した科目に依存する結果であることに留意されたい。なお，前川（2020a, 2020b）は共変量を用いない方法を提案しているが，公民の場合には２科目受験者の数が極端に少ないため，共変量の利用は必須であると考えられる。

また，(3)は，(1)のうち，共変量の全 3 科目に解答した受験者に限るという意味である。例えば，本分析からは，国語・英語 R には受験したが，英語 L を受験しなかった受験者は除外されている。

(4)の weak unconfoundedness の仮定（Imbens, 2000）とは，共変量調整を行う上での必須の統計的仮定である。これは，本研究の文脈でいえば，共変量（国語・英語Ｒ・英語Ｌ）が所与のとき，公民科目選択と公民科目得点は独立であることを要請するものである。別の言い方をすれば，どの公民科目を選択するかの情報は，すべて共変量３科目に含まれているということである。厳密に言えば，現実的に成立していない厳しい仮定であるが，実際にはこれ以上仮定を緩めると共変量調整という技術を用いることが非常に困難になってしまうため，共変量調整を行う上では通常の仮定となっている。しかし，共変量の数を増やしていけば，一般的には，この仮定を満たしやすくなる。上述したように，科目得点の共変量としては大受験者数科目である国語・英語Ｒ・英語Ｌに限定されるが，それ以外の受験者の情報である性別・出身高校などを用

いることもできる。また，Imbens & Rubin (2015)のように，科目得点同士の積の項や科目得点と性別の交互作用項なども共変量として用いることができる。これらの共変量を追加することで，傾向スコア推定時のロジスティック回帰分析の説明率は向上し，傾向スコアを用いることの妥当性はさらに増すと予想され，今後の検討課題としたい。

最後に，(5)に挙げたように，本分析で傾向スコアを推定する際に用いた統計モデルはロジスティック回帰分析であった。ロジスティック回帰分析は，傾向スコアを推定する統計モデルとして最も頻用されている統計モデルである。しかし，傾向スコアを推定する手法は，ロジスティック回帰分析でなくてはいけないということはない。判別分析やニューラルネットワークなどを用いることもできる。一般的には正判別率が高いモデルを用いたほうが良い（星野・前田, 2006）。したがって，本分析の結果は，ロジスティック回帰分析に基づく傾向スコアを用いた共変量調整の結果であることにも留意されたい。

以上，本分析における制約についてまとめた。これらの制約を加味したうえで，結果を吟味する必要がある。統計的手法は，様々な仮定を置いた上で分析されるものである。その仮定を１つ１つ可視化していくと，受け入れ可能な仮定もあれば，受け入れがたい仮定，現状では受け入れざるを得ない仮定などに分かれる。そして，統計分析はいくつかの仮定の集合の下での分析結果であるため，なかなか断定的なことを言うことができないが，こういった分析を積み重ねていくことは肝要である。

現在，共通テストで用いられている得点調整方法は，分位点差縮小法（前川, 2001）である。便利で簡単な手法であるが，受験者集団の学力差を考慮した手法ではない。本研究は，今後，得点調整方法を改善する必要に迫られたときの１つの考え方として示すものである。

注

1) 最終結果は，得点調整が実施・反映された結果であるため，得点調整が反映されていない素得点の中間発表資料を示している。

2) 本論文の見解は，著者らの個人的見解であり，所属組織の公式見解ではありません。

3) 性別・現浪・出身高校なども共変量として指定することができるが，得点以外のステータス変数を学力調整に使用することは過去の同様な研究（大津, 2011; 石岡, 2011; 吉村・荘島, 2004; 荘島ら, 2006, 2007）にも例がなく，またいくつか乗り越えるべき議論があるため（得点以外の情報を調整に使うべきでないという論理的側面と，質的変数なので共変量として安定して使いにくいという技術的側面など），ここでは用いない。

4) 応用上，素点に対して直接IPWを行うことが多いが，Firpo (2007)のように任意の分位点で重みづけることができる。本研究では，得点xが区間tに入るとき 1，そうでないとき 0 であるような 2 値変数fに対して IPW を行っている。これは，相対度数の調整平均を推定していることと同義である。当然であるが，得点を直接調整して求めた平均値と，本分析のように分布を先に調整してから求める平均値は一致する。

参考文献

Austin, P. C. (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, **46**, 399−424.

大学入試センター (2020).「令和３年度大学入学選抜に係る大学入学共通テスト受験案内」大学入試センター (https://www.dnc.ac.jp/albums/abm.php?f=abm00038531.pdf&n=00_令和３年度共通テスト受験案内.pdf) (2021年8月8日).

大学入試センター (2021a).「令和３年度大学入学共通テスト実施結果の概要」大学入試センター (https://www.dnc.ac.jp/albums/abm.php?f=abm00040282.pdf&n=実施結果の概要) (2021年8月8日).

大学入試センター (2021b).「令和３年度大学入学共通テスト公民換算表」大学入試センター (https://www.dnc.ac.jp/albums/abm.php?f=abm00040204.pdf&n=令和3年度大学入学共通テスト（1月16日・17日）公民換算表.pdf) (2021年8月8日).

大学入試センター (2021c). 「令和３年度大学入学共通テスト理科②換算表」大学入試センター (https://www.dnc.ac.jp/albums/abm.php?f=abm00040205.pdf&n=令和3年度大学入学共通テスト（1月16日・17日）理科②換算表.pdf) (2021年8月8日).

大学入試センター (2021d).「令和3年度大学入学共通テスト(1月16日・17日)平均点等一覧(中間集計)」大学入試センター (https://www.dnc.ac.jp/albums/abm.php?f=abm00040200.pdf&n=【HP】中間集計.pdf) (2021年8月8日).

Firpo, S. (2007) Efficient semiparametric estimation of quantile treatment effects. *Econometrika*, **75**, 259–276.

橋本貴充 (2021).「公民と数学の分析」『日本テスト学会第19回大会抄録集』, 30−33.

菱山完・岡田謙介 (2019).「PISA2015における探求型教授法が理科の到達度に与える因果効果の検討」『日本テスト学会誌』

15, 135－148.
星野崇宏 (2009). 『調査観察データの統計科学−因果推論・選択バイアス・データ融合』岩波書店.
星野崇宏・前田忠彦 (2006).「傾向スコアを用いた補正法の有意抽出による標本調査への応用と共変量選択の提案」『統計数理』**54**, 191－206.
Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, **87**, 706–710.
Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.
石岡恒憲 (2011).「Random Forestを用いた欠測データの補完に基づく大学入試センター試験科目間得点差」『応用統計学』**40**, 193－209.
Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, **188**, 250－257.
前川眞一 (2001).「大学入試センター試験における選択科目間の得点調整について」『計測と制御』**40**, 568－571.
前川眞一 (2020a).「加算モデルを用いた選択科目における集団の学力と試験の難易度の分離」『大学入試センター研究開発部リサーチノート』RN-20-06.
前川眞一 (2020b).「成績データから見たセンター試験」大学入試センター『「センター試験」をふり返る』, 164－187. (https://www.dnc.ac.jp/albums/abm.php?f=abm00040328.pdf&n=「センター試験」をふり返る.pdf) (2021年8月8日).
大津起夫 (2004).「潜在変数の区分多項式変換を用いた非線形因子分析」『行動計量学』**21**, 1－15.
大津起夫 (2011).「大学入試センター試験における科目別得点の非線形因子分析による比較」『大学入試センター研究紀要』**40**, 1－23.
斎藤知洋 (2020).「シングルマザーの正規雇用就労と経済水準への影響」『家族社会学研究』**32**, 20－32.
Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41－55.
荘島宏二郎・石塚智一・橋本貴充・大津起夫・前川眞一 (2007).「多重指標モニタリングによる得点調整手続きの試案」『大学入試センター研究紀要』**36**, 53－70.
荘島宏二郎・吉村宰・大津起夫・田栗正章 (2006).「傾向スコアを用いた等百分位法」『大学入試センター研究開発部リサーチノート』RN-06-05.
吉村宰・荘島宏二郎 (2004).「本追モニター調査の結果を利用した傾向スコア加重法による本追試験間の難易度比較」『大学入試センター研究紀要』 **33**, 19－28.