論　文

成果評価基準における正確性について
—フィギュアスケートにおける主観的評価の検証

行武　憲史*
日本大学

藤野　玲於奈**
慶應義塾大学

成果主義の導入が進む中、評価の正確性の重要度は増している。本研究では、人事評価で重視される主観的評価におけるバイアスの存在について、フィギュアスケートの採点データを用いて検証した。より主観的な評価指標である演技構成点と、より客観的な評価指標である技術点のそれぞれについて、滑走グループ間における得点差の有無を RDD（回帰不連続デザイン：Regression Discontinuity Design）により分析した。その結果、能力差を考慮しても、構成点ではグループ間の差が確認され、一方で技術点では確認されなかった。これは、構成点という主観的評価指標には、審判の選手に対する期待によるバイアスが含まれることを示している。

1. はじめに

2015 年 4 月に閣議決定された労働基準法改正案は、「時間ではなく成果で評価される働き方を希望する労働者のニーズに応え、その意欲や能力を十分に発揮できるようにすること」を目的としている。「脱時間給」制度（ホワイトカラー・エグゼンプション）を新設するほか、大企業を中心に 1990 年代半ばから導入が進んできた裁量労働制の対象の拡大などが盛り込まれている。

こうした中、今後重要性を増すと予想される成果主義型の賃金制度については、その是非について様々な分野で活発な議論がなされてきた。例えば、高橋（2004）は経営学の立場から、生産性と金銭評価を直接的に結び付ける成果主義は、仕事内

本論文の審査の過程では、匿名レフェリーの先生方から大変有益なコメントをいただいた。この場を借りて感謝の意を表したい。また、資料整理にあたり、松下駿平氏には多大な協力をいただいた。さらに、浅田義久氏（日本大学経済学部）、川西諭氏（上智大学経済学部）、瀬下博之氏（専修大学商学部）、千木良弘朗氏（東北大学大学院経済学研究科）、直井道生氏（慶應義塾大学経済学部）、中川雅之氏（日本大学経済学部）、森泉陽子氏（神奈川大学経済学部）をはじめ、多くの方々から貴重なコメントをいただいた。日本スケート連盟には、採点方法に関する資料の提供を受けた。記して感謝いたしたい。

* （連絡先住所）〒101-8360 東京都千代田区神田三崎町1-3-2 日本大学経済学部
（E-mail）yukutake.norifumi@nihon-u.ac.jp
** 慶應義塾大学経済学研究科博士前期課程修了（所属は執筆時のもの）

容そのものから効用を得る内発的動機付けを阻害するため[1]、結果として労働意欲を制限し生産性を低下させると論じている。また、阿部（2006）は、従来の職能資格制度ではモラルハザードが発生することを指摘すると同時に、成果主義的賃金制度でもインセンティブの与え方次第ではかえって生産性を低下させるなどの問題点があり、いずれの制度も完全ではないとしている。

樋口（2006）および大湾（2011）は、成果主義的な賃金制度を導入するにあたり、こうした評価基準および評価制度の慎重な設計を求めている。すなわち、評価制度は企業の経営戦略と密接に関係しているため、企業の個別性を踏まえることが重要であるとし、労働者のインセンティブを高めるために、客観的成果指標と主観的評価指標といった複数の指標軸をどのように組み合わせて評価を行えばよいか[2,3]、さらに、どのように運用して処遇に反映させるかについて議論している。大湾（2011）は、効率的な評価制度において、客観的評価と主観的評価は補完的な関係にあるとしている。すなわち、客観的指標は、景気の変動や産業構造の変化など、従業員の努力だけではコントロールできない多くの不確定要素を含むため、従業員がリスク回避型であれば効率的な努力水準を達成できない。そこで、顧客満足度など様々な視点から、従業員の努力水準を評価した主観的指標を組み合わせることにより、効率的な従業員の努力を引き出す。実際、阿部（2006）によると、日本的な成果主義では、客観的な成果指標のみに基づいて人事考課を行う企業はまれである。

一方で、樋口（2006）では、評価システムや評価結果の伝え方など労働者の評価に対する納得感が重要であるとしている。現状、人事評価の仕組みに不満を抱く労働者は少なくなく、それは評価制度の不備や不透明性によるところが大きい[4]。大竹・唐渡（2003）は、こうした労働者の納得感の重要性について実証的な分析を試みた。具体的には、成果主義的な賃金制度の導入が労働者の労働意欲に与えた影響について、労働者・企業の両方に対するアンケート調査データによる分析を行った。その

---

[1] 「内発的動機付け」についての詳細な定義については、Deci（1975）を参照のこと。

[2] 大湾（2011）は、評価指標の 4 つの軸として、客観的指標・主観的指標のほか、企業収益に直結する指標か結果への過程も重視するかというアウトプット指標・インプット指標、目標達成のみを基準とするか評価ランクを用いるかという絶対的指標・相対的指標、集団的業績指標・個人的業績指標を挙げている。

[3] 大湾（2011）は、契約理論上、客観的指標と主観的指標の違いは、契約で合意事項の実施が強制できるかどうかであるとしている。すなわち、労使間の合意に基づいた目標達成について外部機関の確認が容易であれば客観的、そうでなければ主観的であるとしている。

[4] 2015 年に NTT コムリサーチと日本経済新聞社によって実施された『人事評価に関する調査』は、「NTT コムリサーチ」登録モニターのうち、男女 20－50 代の会社員を対象とした。回答者のうち、人事評価の仕組みに 37.8%が不満を示しており、その理由として「評価基準が明確に示されていない」(41.0%)、「評価者の好き嫌いで評価されてしまう」(38.7%)、「評価者が直属の上司しかおらず、評価が一面的」(24.9%)など評価基準の不透明性が上位にある。ただし、人事評価に関する調査には、その評価制度に不満を持つ者が回答しやすい可能性があり、サンプルセレクションの問題に注意する必要がある。

結果によると、成果主義的な賃金制度導入は平均的には労働意欲に影響を与えず、その影響は労働者グループで異なる。特に、現行の人事評価について否定的な考えを持つ労働者は、成果主義の導入によって意欲が低下することが確認され、労働者全体の意欲を高めるためには、労働者に評価基準及び制度を明確に認識させることが重要であるとしている。

そのためには、どのような状況下で労働者が納得感を持つかという議論だけでなく、評価の正確性についての検証が欠かせない。評価の過程に様々な評価バイアスが存在する点は、心理学を中心に多くの研究で指摘されている。バイアスの存在・バイアスへの対処の欠落は、評価の正確性を阻害するために不公平感を招き、従業員の勤務態度などを通じて生産性の低下をもたらすと考えられる。こうしたバイアスは、売上・生産額といった客観評価に比べ、観察しにくい主観的評価においてより高い確率で生じると考えられるため、本研究においては主観的評価指標に焦点を当て、評価におけるバイアスの存在を検証する。

組織内で主観的評価におけるバイアスが認識される場合、労働者はより客観的な評価がなされる業務に努力を傾けるかもしれない（いわゆる、マルチタスク問題）。このとき、最適な労働資源配分からの乖離が生じ、企業の生産性が低下する。客観的な成果指標よりも、主観的評価が人事考課に強く影響している日本の現状では（阿部，2006）、主観的指標におけるバイアスの存在は、成果主義的制度の是非の問題に限らず、報酬制度全体の有効性を揺るがす大きな問題となりうる。

分析にあたっては、企業の人事データを用いることが最善であるが、大湾（2011）が指摘するように、評価バイアスについてのそのような定量分析は多くない。これは、人事データが入手困難なこと、入手しても各評価の個別性が高く、統一した基準のデータが得られにくいことが原因として挙げられる。一方で、スポーツ分野のデータにおいては、客観的なパフォーマンスと審判による主観的評価の組み合わせが比較的容易に観察可能で、かつ主観的評価基準であっても評価項目が明示されることが多く、定量的な分析に適しているため、成果評価の分析に多く用いられている。本研究においても、フィギュアスケートの国際大会における成績データを用い、そのデータの特性を利用して評価バイアスの検証を行う。フィギュアスケートの得点のうち、演技構成点（かつての芸術点。以下、構成点と記す）は表現力に対する測度であり、評価基準は相対的に主観性が強い。そのため、審判は構成点の評価に際して特定の選手に対する期待を評価に反映させる余地（バイアス）が生じる。一方で、技術点は、選手が実行した演技をビデオ判定などによって一定の客観的基準を用いて評価されるため、こうしたバイアスが生じにくいと考えられる。

本研究で対象とするのは、演技グループごとに生じる期待バイアスである。フィギュアスケートにおいては、整氷作業や選手に公平なウォーミングアップの時間を与えるため、６人ごとのグループに分けられる。グループ分けは成績順に行われるため、後半のグループほど能力の高い選手が集まる可能性が高くなる。このとき「能力の高い選手が集まるグループだからきっといい演技をするに違いない」といった期待に基づく評価がなされれば、ほぼ能力の等しい選手であっても振り分けられたグループによって、評価に大きな違いが生じる可能性がある。分析には、グループ分けによって生じる不連続性を利用して RDD（回帰不連続デザイン：Regression Discontinuity Design）を用いた。05 年から 15 年までの世界選手権とオリンピック３大会の男女シングル、14 大会 28 競技のデータを用いて、（より主観的な）構成点と（より客観的な）技術点を比較して分析を行った結果、構成点においてのみ期待バイアスの存在が確認された。さらに、バイアスの影響を除いた得点を推計した結果、特定の選手においては最終順位が変わる可能性が示された。この点は、主観的評価が客観的評価に比べバイアスを生じさせる可能性が高く、評価基準の設計・運用に当たってはこうしたバイアスを考慮した制度が必要である点を示唆する。

本研究の構成は以下の通りである。次節においては、検証の過程をより明確にするため、フィギュアスケートの採点方法について説明する。第３節では、先行研究をもとに主観的評価に基づくバイアス、特に本研究で対象とする期待バイアスについて整理する。その次の節では実証モデルを提示し、第５節においては、分析に用いたデータを紹介する。第６節は分析結果の紹介とバイアスを考慮したシミュレーションを行う。第７節は結論である。

2. フィギュアスケートの採点方法

本節においては、検証の過程をより明確にするため、フィギュアスケートの採点方法について解説する。フィギュアスケートの採点システムは、02 年のソルトレイクオリンピックでの審判スキャンダルの発覚により、04 年に採点方式が大幅に変更された。最大の変更点は、旧システム（6.0 システム）に比べて、新システムにおいて評価の相対性がかなり弱くなった点である。すなわち、旧システムでは、技術点・芸術点のいずれも、審判が選手それぞれについて６点満点で相対的に採点し、最終順位は各審判が出す順位を得点化した順位点で競われた。一方で、新システムでは、下記に示すように、技術点・構成点共に、各選手に絶対的な評価が与えられ、その合計点によって順位が決定される。以下では、14／15 年シーズンのルールをベース

に採点システムを紹介する[5]。

2.1 フィギュアスケート競技会の概要

フィギュアスケートのオリンピック及び世界選手権における競技は、おおむね２日間にわたり開かれる。大会の１日目は、ショートプログラム(以下、SP と記す)が開催される。SP の特徴は、演技時間が２分 50 秒以内と短い点と、ジャンプやスピンなどの７個の必須要素があることである。SP での滑走順は、直前の国際スケート連盟(International Skating Union、以下 ISU)の世界ランキング(ISU World Standings)により決定される。２日目は、フリースケーティング(以下、FS と記す)が開催される。FS においては、演技時間が男子は４分 20－40 秒、女子は３分 50 秒－４分 10 秒とそれぞれ長く、男子は 13 個、女子は 12 個までの要素を組み入れた演技から成る。滑走順は、前日の SP の成績により決定される。

フィギュアスケートの得点は SP・FS のいずれも、技術点・構成点・減点の３種類から成る。これら３種類の合計得点をそれぞれ SP・FS の得点とし、SP・FS の得点を合計したものが、その大会の総得点となる((1)式)。

$$Total\ Score = SP\ Score + FS\ Score \tag{1}$$

ISU が主催する大会の場合、全ての審判が ISU 会長によって任命される。判定は大きく分けて、３種類の審判によって行われる。イベントレフェリーは、競技会と審判団を監督する。技術審判は、技の基礎点(Base Value)の判定を行い、演技審判は、GOE（Grade of Execution：各要素の出来栄え）と構成点(Program Component Score)を判定する。

2.2 フィギュアスケートの各得点と採点方法

技術点(Technical Element Score)とは、選手が実行した各規定要素に与えられる得点の合計である。技術点は、基礎点と GOE から成る。基礎点は技術審判が判定し、GOE は演技審判が判定する。それぞれの得点は、選手の演技終了後に即座に判定され、旧システムのように選手間の相対評価は行われない。得点の構成は以下の式で表すことができる。

$$SP(FS)Score = Base + GOE + PCS - Ded \tag{2}$$

基礎点（Base）は、実行された技の基礎的な評価である。評価のポイントは３点あり、１つ目は要素の入り方であり、各要素の開始時の動作(ジャンプの踏切りなど)

[5] 採点に関する説明の主な部分は、International Skating Union（2014）を参照している。

を判定する。２つ目は、ジャンプ要素で選手が実行した回転数の判定である。３つ目は、ジャンプ要素以外の全ての要素を判定するレベル基準である。基礎点の採点は、テクニカル・コントローラー(１名)、テクニカル・スペシャリスト、アシスタント・テクニカル・スペシャリスト(１－２名)の最多で３名の技術審判が担当する。基礎点の採点は演技直後に行われ、映像を用いて各ポイントの細かなチェックも行われるため、比較的明確な基準を持つ。従って、基礎点の評価にはバイアスが少ないと考えられる。

GOE は、演技審判によって判定される、各要素の出来栄えに対する加点・減点を示す。０を基準とし、－３から３までの７段階で評価される。演技審判は９人で構成され、審査手順は、（１） ９人の審判が GOE を判定する。（２） ９人の審判から事前の抽選に基づいて２人を除外する（計７名が残る)。（３） ７人の審判の中で、最大値と最小値を除外する（計５名が残る)。（４）残された５人の審判による点数の平均点が最終的なGOEの点数となる。GOEの場合、加点基準は少し抽象的である一方で、減点基準は転倒・回転不足など、具体的に定められており客観性が高い。

構成点（PCS）は、GOE と同じ演技審判によって、５項目（表１）についてそれぞれ 0.25 単位、10 点満点で採点される。各項目に設定された係数を乗じ、各項目の点数と係数の積の合計が最終的な構成点となる。構成点の審査手順は GOE の場合とほぼ同様で、項目ごとに９人から選別された５人のレフェリーによる点数の平均を用いて計算される。構成点の５項目は、技術点に比べると基準が不明確であるため、審判の主観的判断に依存する部分が存在する。例えば、表１の最後にある、「音楽の解釈」という項目にある「音楽のニュアンスを反映した細部の表現」など、客観的な基準を設定することが明らかに難しい要素が含まれている。

最後に、減点(Deduction：Ded)は、転倒・演技時間超過、その他違反行為などによる減点である。必ずしも起きる事象ではないこと、映像などの利用によりほぼ完全に客観的基準によって減点されることから今回は分析の対象としていない。

表１に示されるように、基礎点は採点基準がより客観的である。GOE は減点部分に関する採点基準の客観性は高いのに対し、加点部分に関する採点基準はそうではない可能性がある。構成点は採点基準が明確ではなく、審判の主観的判断に依存する。

3. 評価バイアスの種類とグループ間の不連続性の要因

3.1 フィギュアスケートの採点における評価バイアス

Kahneman(2012)は、人間が犯すエラーの中でも、特定の状況で繰り返し起きるシステマティックなエラーをバイアスと呼んでいる。Morgan and Rotthoff (2014)は、

表１ 演技構成点の５項目における採点基準

項目	演技構成点の評価で考慮される要素	
スケート技術 Skateing Skills	· Balance and rhythmic knee action and precision of foot placement · Flow and effortless glide · Cleanness and sureness of deep edges, steps and turns · Power/energy and acceleration · Mastery of multi directional skating · Mastery of one foot skating	· バランス、リズミカルな膝の動き、足運びの正確さ · 滑りの流れ、効率的な滑走 · 深いエッジ、ステップ、およびターンの明確さと安定性 · パワー／エネルギー／加速 · あらゆる方向へのスケーティングの熟練度 · 片足スケーティングの熟練度
要素のつなぎ Transitions/Linking Footwork and Movements	· Variety · Difficulty · Intricacy · Quality (including unison in Pair Skating)	· 多様性 · 難易度 · 複雑さ · 質（ペアスケーティングにおけるユニゾンを含む）
動作／身のこなし Performance／Execution	· Physical, emotion, and intellectual involvement · Carriage · Style and individuality/personality · Clarity of movement · Variety and contrast · Projection	· 身体、感情、知性の表現 · 身のこなし · スタイル、個性/パーソナリティ · 動作の明確さ · 動作の多様性とコントラスト · 投影
振り付け／構成 Choreography／Composition	· Purpose (idea, concept, vision, mood) · Proportion (equal weight of parts) · Unity (purposeful threading of all movements) · Utilization of personal and public space · Pattern and ice coverage · Phrasing and form (movements and parts structured to match the phrasing of the music) · Originality of purpose, movement and design	· 目的（アイデア、コンセプト、ビジョン、ムード） · 調和あるプログラム構成 · 統一性（すべての動作が目的をもってつながっているか） · スケーターと観衆の表現空間の利用 · プログラム・パターンの独創性、氷面の十分な利用 · フレーズとフォーム（音楽のフレーズに合わせた動作や各要素の構成） · 目的、動作、デザインの独創性
音楽の解釈／タイミング Interpretation of the Music/Timing	· Effortless movement in time to the music (timing) · Expression of the music's style, character and rhythm · Use of finesse to reflect the nuances of the music	· 音楽にしっかりと合った軽快な動作（タイミング） · 音楽のスタイル、特色、リズムの表現 · 音楽のニュアンスを反映した細部の表現

注：International Skating Union (2014) より。日本語訳は筆者による。

フィギュアスケートなどの採点行為を伴う競技会における様々な種類のバイアスを、期待（参照）バイアス、国籍（内集団）バイアス、順序バイアス、難易度バイアスという４種類に分類している[6]。

期待バイアスとは、選手の演技について、演技そのものではなく期待に基づいて選手の評価を行う傾向を示す。つまり、明確な判断基準が無い場合、現在の演技に対する評価を、演技以前の選手の名声・情報などの知識、あるいは同一の選手に対して過去に行った評価に基づいて行うため発生するバイアスである。本稿の期待バイアスは、Tversky and Kahneman(1974)および Kahneman(2012)において紹介されているヒューリスティックスという考え方に基づく。ヒューリスティックスとは、判断を行う際の情報処理経路のショートカットという直感的な判断方法と定義される。具体的には選手あるいは集団の演技について、実際の演技内容を純粋に評価するの

[6] ただし、Morgan and Rottoff(2014)では、バイアスについての明確な定義づけはなされていない。そのため、必ずしも Kahneman(2012)によって定義されたバイアスとは一致しない。

ではなく、頭に思い浮かびやすい情報、思い浮かべる回数などを参照して評価することと考えられる。

内集団バイアスは、審判と選手が所属する集団（国籍、人種、出身地など）が同じ場合に発生するバイアスである。例えば、国籍バイアスを研究しているものとして、Campbell and Galbraith (1996)と Zitzewitz (2014)がある。これらの論文では、オリンピックのフィギュアスケート競技において、審判と選手の国籍が同一であると高めに採点されることを示した[7]。

順序バイアスは、演技の順序の違いによって発生するバイアスである。Bruine de Bruin (2005)は、フィギュアスケートや歌唱コンテストにおいて、初めの演技者については情報が少ないために極端な採点が行われ、後に演技をするほど得点が上がる点を検証した。

最後に、難易度バイアスは、難易度に関する評価と達成度に関する評価が独立に決定されるべき場合に、実際には同じ達成度であったとしても、より難易度の高い演技について達成度の評価が高くなるバイアスである。Morgan and Rotthoff (2014)は、体操競技において、期待や順序、国籍バイアスをコントロールした上で難易度バイアスの存在を示し、その結果として、選手はより難易度の高い演技を行うインセンティブを持つ点を確認した。

評価バイアスについては、人種などの内集団バイアスに焦点があてられる研究が多かった。Morgan and Rotthoff (2014)と同様の考え方に基づく期待バイアスについての数少ない先行研究としては、Findlay and Ste-Marie (2004)がある。彼らは、カナダの２地域（ケベック州とオンタリオ州）のフィギュアスケートにおける期待バイアスの存在を実証した。当論文が執筆された時点における評価システムは、旧採点システム（6.0 システム）であり、技術点、芸術点がそれぞれ相対的評価に基づいて採点され、最終順位も各審判の順位を得点化した順位点で競われた。分析の結果、最終的な順位点・技術点において期待バイアスが有意に確認され、芸術点にも有意でないものの同様の傾向が確認された。すなわち、審判が事前に選手の評判・情報を知ることにより、その選手に対する過度な肯定的評価が導かれる。一方で、演技の減点など絶対基準に近い評価では同様のバイアスは確認されていない。

[7] 内集団バイアスについては、フィギュアスケート以外の分野でも多くの先行研究が存在する。例えば、Elvira and Town (2001) は、企業のデータを用いた数少ない分析である。米国における販売社員の人種と上司の人種が人事評価結果に与える影響を分析し、上司と部下の人種が異なる際に、評価が有意に低下することを確認した。Price and Wolfers (2010) は、米国プロバスケットボールにおいて、選手と審判の人種が同じ場合にファウルの頻度に差異が発生するかを検証し、人種バイアスを確認した。Parsons et al. (2011) は、米国大リーグにおいて、投手と審判の人種が一致する場合にストライクと判定される確率が高くなる点を示し、投手の方もボール寄りのコースに投球する傾向を確認した。

本研究では、上述のバイアスのうち、フィギュアスケートの滑走グループが成績によって外生的に割り振られる特徴を利用して、新システムにおける期待バイアスについて検証を行う。様々なバイアスについての存在を確認し、潜在的な大きさを測ることは、それを補正する対策を考慮する上で重要である。フィギュアスケートに限らずスポーツにおける審査は、目の前に実行された演技に対し同一の視点でなされるべきであると考えられるが、実際は過去の自身の審査経験や、事前の選手情報に影響を受ける可能性がある[8]。このとき、審判の審査基準に影響する要因には、事象の思い出しやすさ（検索容易性）に基づく利用可能ヒューリスティックスのみでなく[9]、Kahneman(2012)で議論された、思い出した数を根拠として判断する多数性（numerosity） ヒューリスティックス[10]、思い出した内容に基づくアンカリング、代表性ヒューリスティックスも考えられる。本研究においては、審判による期待バイアスにそれらの要因が全て含まれると定義する。

3.2 演技グループ間の不連続性の要因

フィギュアスケートの世界選手権・オリンピックでは、SP および FS の両方において、滑走順が後のグループになるほど実力の高い選手が登場する。すなわち、SP の滑走順は直前の世界ランキング、FS の滑走順は SP の順位によりそれぞれ決定され、上位の選手が後のグループへと振り分けられ、基本的に６人が１グループとして演技をする。このとき、グループに対する期待バイアスが発生し、グループ間に得点の不連続性が生じる可能性がある。一方で、グループ内の滑走順は、ランキングや順位に関係なくランダムに割り振られるため、グループ内部にはこうした不連続性は発生しないと考えられる。

後半のグループは実際に有力な選手が多く集まりやすいため、審判は過去において後半のグループにいる選手に高い評価を与えた経験が多い。その意味で、後のグループになるほど、高い点数を付けた経験の思い出しやすさや思い出せる回数が多くなると考えられる。また、過去の採点経験はアンカー、すなわち参照点として作用し、最終グループに残る選手は表現の高い演技をするというイメージ（代表性）に影響される可能性も高くなるだろう。上記の要因は、全て同じ方向、すなわち、過去の優れた演技が今回の演技についても良いと期待させると考えられる。本研究

[8] 実際、荒川（2013）では、「例えばずっと下の順位にいた選手が、ある日突然に素晴らしい演技をしたとしても、そのとき１回だけではサプライズの結果が出ないのが普通です」と記されており、フィギュアスケートの採点が必ずしも大会ごとに独立したものになっていない可能性を示唆している。
[9] 検索容易性については、森（2008）が詳しい。
[10] 多数性ヒューリスティックの詳細については、Ofir et. al（2008）を参照のこと。

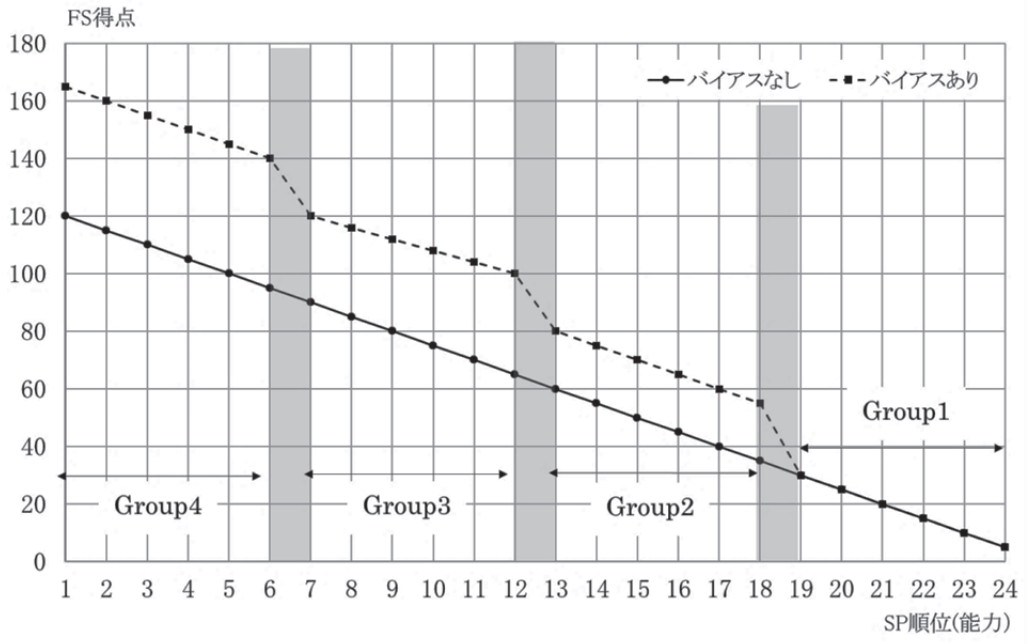
は各要因の識別は行わず、審判が目前の演技とは別の要因に影響を受け、審査結果が歪められる点を確認する。従って、それらの要因を合わせて、グループ間に生じる期待バイアスと定義する。

グループ分けは、SP が世界ランキング、FS は SP の順位によって行われるため、バイアスが存在しない場合にも、選手の能力自体によってグループ間の平均得点には差が生じる。しかし、能力の高いグループほど高い期待によるバイアスの恩恵を受けるならば、グループ間の平均得点差は能力以上に拡大すると考えられる。図１は、バイアスが存在する場合としない場合、それぞれについて選手の能力と得点との関係のイメージを表す。横軸に SP の順位（評価が正確な場合における選手の能力、ランクが小さいほど高い）、縦軸には FS の得点（score）をとっている。図１の実線はバイアスが存在しない場合であり、選手の得点と能力は連続的な関係にある。破線はバイアスが存在する場合であり、バイアスによってグループの境界（６位と７位、12 位と 13 位、18 位と 19 位）において不連続性が生じる。以下の実証分析では、選手の能力をコントロールした後でも、こうした平均得点のグループ差が確認されるか否か、すなわちバイアスが存在するか否かを検証する。

一方で、個々の選手に対する期待バイアスも存在すると考えられる。後のグループの選手は、実際の能力および過去の実績が高く、審判が選手に対する情報を多く有すると考えられる。このとき、グループが後になるほど、個々の選手採点でも高い期待バイアスが発生すると考えられる。逆に、こうした事前の期待が負のバイアス、すなわち当初の期待が大き過ぎる場合に、期待水準以上の演技をしなければ良い評価が得られないバイアスが存在するかもしれない。この「期待外れ」によるバイアスは、実際の演技と、ヒューリスティックスが想定する事前情報との比較によって初めて生じる。採点は、個々の演技終了直後に行われるため、グループ一括ではなく選手ごとに行われる。従って、個々の期待バイアスの影響は、平均的には本研究で想定するグループ間の不連続性には表れないと考えられる。さらに、こうした個々の期待バイアスについては、個々の SP の総得点およびシーズンベストといった能力の代理変数によってコントロールされている。ただし、これらの変数の係数において、採点に対する実際の能力の効果と期待バイアスの効果を識別するのは難しい。

また、グループごとに平均的な得点差を発生させる要因として、選手の本来の実力および、審判による採点のバイアスのほかに、グループ内における演技者のパフォーマンスが相乗（連鎖）的に他の演技者に影響するような効果も考慮する必要がある（ピア効果）。すなわち、前の選手が良い演技をした場合に、対抗意識から優れ

図 1 バイアスの有無別 得点と能力の関係(イメージ図)



た演技が引き出されるケース、あるいは前の選手の失敗演技の影響を受けて質の悪い演技を行う場合である。Ikegami and Ganesh(2014)は実験により、ダーツの熟練者が初心者のプレーを観察することによって、自身の成績が下がる点を示し、他人の行動が自身の行動を歪めることを示した。その上で、最適なパフォーマンスを保つ上では他の選手を観察すべきでないとしている。また、Yamane and Hayashi (2015)は、競泳のデータを用いて、隣のレーンの選手の存在がパフォーマンスに影響する点を示した。特に自分よりも遅い選手に追いかけられる時に成績が向上する。

一方で、他の選手の演技の影響には、逆の方向の影響も存在する。すなわち、前の選手の優れた演技がプレッシャーとなって自身の演技の質が低下する、いわば負のピア効果である。06 年トリノオリンピックにおいて金メダルを獲得した荒川静香氏は、演技前にリンクに背を向けヘッドホンをしていた。雑誌のインタビューに対して、「前の人の結果を知るとうまくやらなきゃと思ってしまうタイプで、(中略)人の演技に振り回されてはいけないと意識していた」と述べている(荒川, 2006)。ただし、この負のピア効果は、良い演技をした選手の影響が他の選手にマイナスに、悪い演技をした選手の影響が他の選手にプラスに影響し、グループ全体としては相殺されるため、グループの平均的評価を一方向に変化させない。そのため、グループの境界における不連続性を生む要因とはならないと考えられる。

ピア効果に加えて、上位グループに属することによってモチベーションが高まる

発奮効果も考えられる。すなわち、本来グループ境界付近の実力を持つ選手が、SPの成績によって上位グループに振り分けられたとき、より上位進出へのモチベーションが高まり、平均以上の演技が引き出される可能性がある。逆に、下位グループに属する場合は、モチベーションが下がり凡庸な演技を行うかもしれない。その場合、境界付近で同様の実力の選手であったとしても、上位グループに属する選手の得点は平均的に高まり、逆に下位グループの選手の得点は平均的に低くなり、グループの境界で不連続性が発生する。

期待バイアス、ピア効果、発奮効果の識別には、基礎点（技術点の一部）と構成点における採点基準の客観性の違いが役に立つ。構成点と技術点は、評価基準の客観性に違いがあるものの、その要素は密接に関係すると考えられるため、ピア効果や発奮効果が選手のパフォーマンスを向上させるならば、その影響は両方に及ぶと考えられる[11]。Yamane and Hayashi (2015)は、ピア効果が競泳におけるタイムというシンプルな身体的パフォーマンスを向上させる点を示している。しかし、フィギュアスケートにおいて身体要素の大きい技術点のみがピア効果・発奮効果の影響を受ける場合は考えにくく、グループ間の不連続性は、基礎点、GOE、構成点の全てにおいて発生するだろう。一方、採点基準が相対的に明確でない構成点にのみ不連続性が確認されれば、審判の主観によるバイアスが存在する可能性が高くなる。

ただし、ピア効果・発奮効果が表現面のみに現れる状況を完全に排除することは難しいため、期待バイアスとピア効果・発奮効果を識別する追加的な分析も行った。

4. 実証モデル

今回の分析に当たっては、以下の(3)式のような評価関数をベースに考える。

$$S_i = \alpha + \sum_{j=2}^{4} \beta_j D_{ij} + \sum_{k=1}^{K} \gamma_k X_{ik} + u_i \qquad (3)$$

被説明変数であるS_iは演技者iの対象スコアを表し、基礎点、GOE、構成点を考慮する。D_{ij}はグループダミーを表しており、添え字jがグループ$(j = 2,3,4)$を表し、第1

[11] 野口（2014）によれば、構成点に含まれる「5項目すべてに共通するのは「ムーブメント（動作)」であるとされる。例えば、一般的に考えられる「演技力」にあたる「動作/身のこなし」について、審判は、身のこなし、音楽を捉えた感情表現、そして選手独自の世界観などを評価する。もし技術点であるジャンプでミスが多く、表現がぶつ切りになるようであれば、当然ながら一貫した世界観が崩れてしまうのでこの項目は低く評価される。この場合、ピア効果や発奮効果によってジャンプのパフォーマンスが上がる場合、同時に「動作/身のこなし」の得点も高くなると考えられる。

グループを基準としている。X_{ik}は、個人属性を表し選手の能力・大会ごとの得点傾向などを含む。最後に、u_iは誤差項である。また、α、β_{j}、γ_{k}は各パラメータを表す。添え字は、iが選手個人の大会ごとの演技、kがk番目のその他の説明変数をそれぞれ表す（$k = 1, \ldots, K$）。

FS のグループは、SP の成績に基づき６人ごとの４グループに分けられるため、選手の能力と高い相関がある。(3)式においては、説明変数によって、選手の能力が完全にコントロールされなければ、内生性の問題が発生する。その問題を解消するには、グループ分けが完全にランダムに行われる必要がある。そこで、本研究では RDD 法を用いてこの問題に対処する[12]。グループ分けは６人ずつなので、SP の成績が６位と７位、12 位と 13 位、18 位と 19 位にそれぞれ閾値が存在する。この前後であれば、選手間の能力差は非常に小さいと考えられる。よって、サンプルのうち閾値近辺のデータのみを抽出して検証すれば、能力の近い選手同士の比較になり、擬似的にランダム化した状況に近づく。このサブサンプルにおいて、上位グループの最下位の選手と１つ下のグループの最上位の選手を比較して、能力とは関係ない、グループ間の違いによる FS の得点の違いを確認できる。具体的には、以下のような回帰式を定義して推定を行う。

$$S_i = \alpha + \beta D_i + \gamma SPC_i + \delta SBF_i + \sum_{t=2}^{14} \eta_t TD_{ti} + \varepsilon_i \qquad (4)$$

ここでD_iは、上位グループダミーを表し、それぞれのグループの最下位、すなわち６位、12 位、18 位であれば、１とするダミー変数である。ただし、05、06 年の世界選手権においては第１・２グループが分割されていないため、第 18 位と第 19 位のサンプルは用いない。また、(3) 式のモデルと異なり、グループごとの閾値の差を個別に推定するのではなく、各グループの差の平均をまとめてβとして推定する。

一般に、RDD では閾値が共変量の関数になっており、閾値前後の能力の違いを厳密にコントロールするため共変量を、さらに非線形性を考慮してその高次項をモデルに含める。ここでは、閾値を決める共変量が SP の順位であり序数のため、選手の能力の差を厳密には表さない。そこで、SP の順位のベースとなる SP のスコアを共変量として用いる（SPC_i）。また、非線形性を考慮した高次の項については、実証分析で有意な結果が得らなかったため、実際には１次項のみ用いた。

さらに、SP のスコアは、オリンピックや世界選手権での一度のみの試技によるた

[12] RDD 推定法についての詳細な説明は、Imbens and Wooldridge (2009)や Angrist and Pischke (2009) などを参照のこと。

め、プレッシャーによって実力を発揮できないなど、厳密な選手の能力を表さない可能性がある。そこで、その大会直前までの各個人の FS のシーズンベストの値を加えている（SBF_i）。オリンピックおよび世界選手権は各シーズンの終盤に行われるため、シーズンベストは当年における選手の能力を反映したものと考えられる。より具体的には、各個人の表現および技術に関する能力を正確に捉えるため、大会直前におけるシーズンベストを構成点と技術点に分け、被説明変数が構成点のモデルについてはシーズンベストの構成点を、被説明変数が基礎点と GOE のモデルについてはシーズンベストの技術点を説明変数として導入している。加えて、シーズン間でのマイナーなルール変更や競技レベル向上による大会間の得点傾向の違いをコントロールするため、大会ダミー（TD_{ti}）を含めている。なお、男女間で採点の方法が異なるため、男女別に分析を行う。変数の詳しい定義については表２で示した。

5. データ

データは、世界選手権 11 大会(05－15 年)と冬季オリンピック３大会（06 年トリノ大会、10 年バンクーバー大会、14 年ソチ大会）のシングル男女を対象としている。世界選手権やオリンピックでは、約 30 名の選手が SP を演技し、上位 24 名が FS に進む。また、得点などの各大会の情報は、基本的に ISU の公式ホームページ（http://www.isu.org/en/home）より入手可能である[13]。

表２ 変数の定義

変数名	定義
上位グループダミー	フリースケーティング時のグループ内で、ショートプログラムのスコアが最下位であれば、1を取るダミー変数。つまり、ショートプログラムの順位が、6位、12位、18位の時に1となる変数。
基礎点(FS)	フリースケーティングにおける選手ごとの基礎点
GOE(FS)	フリースケーティングにおける選手ごとのGOE
構成点(FS)	フリースケーティングにおける選手ごとの構成点
総得点(SP)	ショートプログラムにおける選手ごとの総得点
SBの構成点	大会直前のFSにおけるシーズンベストの構成点
SBの技術点	大会直前のFSにおけるシーズンベストの技術点
ピア効果	自分が属するグループ内の他選手のシーズンベスト総得点の平均値から、自分のシーズンベスト総得点を引いた値。

[13] シーズンベストの情報については、日本選手権のように国際大会ではない場合などは、ISU から入手できないため、各大会などのホームページから情報を収集した。収集先の情報については筆者に問い合わせ願いたい。

推定に用いたデータの基礎統計量を表３に示した。標本サイズは、男子が 332、女子が 329 である。FS を棄権した選手などを欠損値とした後、グループの閾値前後にある標本は、男子 80（上位 40、下位 40）、女子 80（上位 40、下位 40）となった。

表３　基礎統計量

男子				
変数名	全体	上位グループ	下位グループ	平均の差
基礎点(FS)	66.79	68.01	66.58	1.43
	(8.46)	(6.88)	(7.27)	[0.90]
GOE(FS)	0.67	1.31	0.70	0.61
	(5.03)	(4.54)	(5.30)	[0.56]
構成点(FS)	68.45	69.84	67.33	2.51
	(10.23)	(8.19)	(9.38)	[1.27]
総得点(SP)	71.29	72.12	70.87	1.25
	(9.84)	(7.02)	(6.80)	[0.81]
シーズンベスト(FS)	140.92	140.17	141.10	−0.93
	(19.79)	(16.36)	(18.10)	[−0.24]
SBの技術点	−	70.87	71.30	−0.43
		(8.90)	(10.35)	[0.20]
SBの構成点	−	69.62	70.23	−0.60
		(8.83)	(8.91)	[−0.30]
ピア効果	−	7.82	−5.96	13.78***
		(10.14)	(13.76)	[5.10]
標本サイズ	332	40	40	

女子				
変数名	全体	上位グループ	下位グループ	平均の差
基礎点(FS)	49.38	49.25	47.87	1.37
	(6.69)	(6.71)	(6.61)	[0.92]
GOE(FS)	0.74	1.09	0.11	0.98
	(4.33)	(3.60)	(3.73)	[1.19]
構成点(FS)	51.18	52.18	49.50	2.68
	(9.17)	(7.09)	(8.05)	[1.58]
総得点(SP)	55.24	55.68	54.61	1.08
	(7.81)	(5.19)	(5.03)	[0.94]
シーズンベスト(FS)	105.68	104.84	106.79	−1.95
	(17.59)	(13.72)	(16.95)	[−0.56]
SBの技術点	−	52.15	54.17	−2.02
		(8.35)	(9.46)	[−1.01]
SBの構成点	−	52.77	52.82	−0.05
		(6.83)	(8.32)	[−0.03]
ピア効果	−	7.66	−6.18	13.84***
		(12.86)	(13.42)	[4.71]
標本サイズ	329	40	40	

注１）上位グループは SP 順位の６位、12 位、18 位の選手を、下位グループは同７位、13 位、19 位の選手をそれぞれ表す。
注２）標本サイズ以外の表内数値は標本平均、丸かっこ内は標本標準偏差、角かっこ内は t 値を指す。
注３）***は１％水準で有意であることを表す。

表３の右側の列においては、上位・下位グループの間について、各変数の平均値の差についてｔ検定を行った。それによると、ピア効果以外の変数において有意な差は観測されない。しかし、構成点については有意水準 10%に近い値を得ている。また、SP 得点やシーズンベストには、少なくとも閾値付近において統計的に有意な能力差が無い。ピア効果は、追加的な分析に用いるもので詳細については後述する。

図２は SP の順位ごとに各大会の FS の基礎点､構成点の平均を取ったものである。男子は基礎点、構成点共に明確な不連続性が確認できない。女子の基礎点においては境界における不連続性が不明確な一方、構成点についてはグループ１・２、グループ２・３の間に不連続な関係がみられる。不連続性はグループの境界以外においてもみられるため、グラフから期待バイアスの有無を判断することは難しい。次節の RDD 分析において、能力をコントロールした上で期待バイアスの検証を行う。

6. 推定結果

6.1 推定結果（RDD）

推定結果は男子を表４上部に、女子を表４下部にそれぞれ示した。それぞれ、左側に不連続点をグループ境界にした場合、中央に不連続点をグループ内に想定した場合の推定結果を示した。被説明変数は左から FS の構成点、基礎点、GOE である。

図２　FS 得点と SP 順位の関係（各大会を通じた平均値）

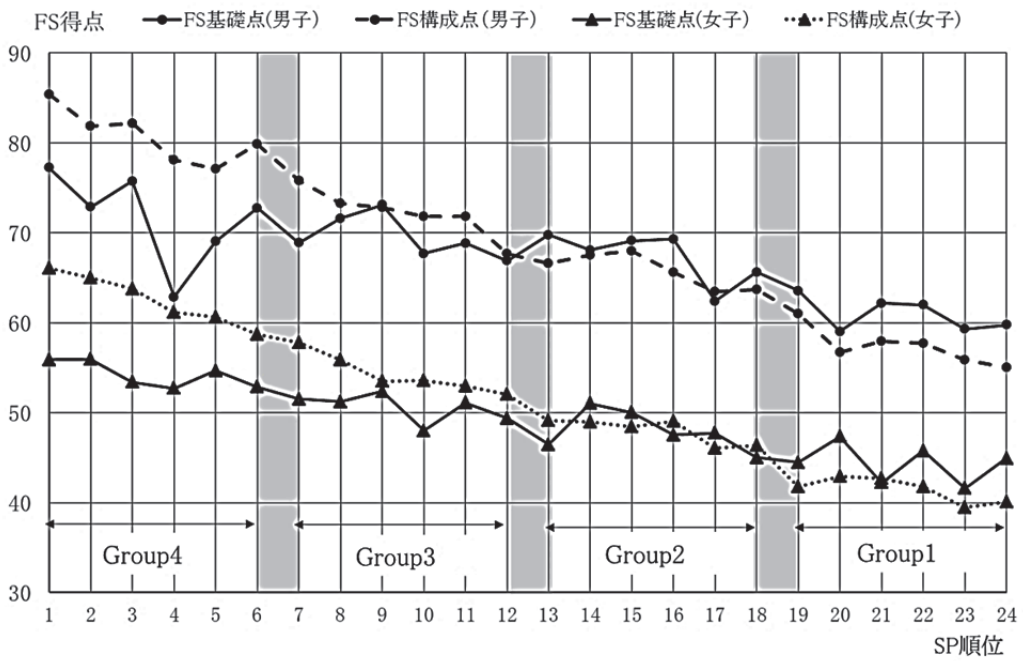


表４ 推定結果

男子

	グループ境界			グループ内:プラシーボテスト			グループ境界:ピア効果
	構成点(FS)	基礎点(FS)	GOE(FS)	構成点(FS)	基礎点(FS)	GOE(FS)	構成点(FS)
上位グループ	2.001***	0.980	0.283	-0.466	-1.403	-0.336	2.332***
	(0.693)	(1.415)	(0.912)	(0.508)	(1.084)	(0.635)	(0.812)
総得点(SP)	0.651***	0.381**	0.296***	0.677***	0.284**	0.271***	0.684***
	(0.065)	(0.146)	(0.086)	(0.050)	(0.114)	(0.060)	(0.081)
SB の構成点/SB の技術点	0.516***	0.074	0.093	0.484***	0.323***	0.048	0.474***
	(0.064)	(0.124)	(0.069)	(0.041)	(0.097)	(0.047)	(0.094)
ピア効果	―	―	―	―	―	―	-0.029
							(0.038)
定数項	-16.504***	33.186***	-28.810***	-13.409***	24.651***	-24.031***	-16.014***
	(4.035)	(9.218)	(6.014)	(3.124)	(6.879)	(3.796)	(4.179)
大会ダミー	Yes	Yes	Yes	Yes	Yes	Yes	Yes
決定係数	0.911	0.392	0.449	0.881	0.344	0.368	0.911
標本サイズ		80			159		80

女子

	グループ境界			グループ内:プラシーボテスト			グループ境界:ピア効果
	構成点(FS)	基礎点(FS)	GOE(FS)	構成点(FS)	基礎点(FS)	GOE(FS)	構成点(FS)
上位グループ	1.912***	1.277	0.939	-0.671	-1.103	-0.551	1.730*
	(0.643)	(1.231)	(0.711)	(0.455)	(0.798)	(0.512)	(0.897)
総得点(SP)	0.737***	0.519***	0.206**	0.733***	0.323***	0.230***	0.704***
	(0.114)	(0.135)	(0.081)	(0.073)	(0.107)	(0.061)	(0.139)
SB の構成点/SB の技術点	0.469***	0.229***	0.090*	0.466***	0.304***	0.105***	0.501***
	(0.076)	(0.071)	(0.052)	(0.047)	(0.052)	(0.037)	(0.117)
ピア効果	―	―	―	―	―	―	0.016
							(0.045)
定数項	-13.606***	6.787	-16.088***	-12.110***	14.003**	-16.900***	-13.373***
	(4.302)	(7.827)	(4.175)	(2.942)	(5.438)	(2.762)	(4.312)
大会ダミー	Yes	Yes	Yes	Yes	Yes	Yes	Yes
決定係数	0.891	0.507	0.465	0.871	0.430	0.351	0.892
標本サイズ		80			158		80

注１）それぞれ、***は 1%水準、**は 5%水準、*は 10%水準で有意であることを表す。
注２）グループ境界は、滑走グループの境界を不連続点とした場合を指す。すなわち、SP 順位の 6 位と 7 位、12 位と 13 位、18 位と 19 位の間を不連続点とする。グループ内はいわゆるプラシーボテストで、SP 順位の 5 位と 6 位、7 位と 8 位、11 位と 12 位、13 位と 14 位、17 位と 18 位、19 位と 20 位の間を不連続点としている。
注３）05、06 年の世界選手権では、第 1、2 グループは SP の成績に関わらず、同じプールから抽選をして順番を決めた。従って、両大会の 18 位と 19 位はサンプルから除外した。
注４）SB については、構成点モデルは SB の構成点(FS)、基礎点(FS)・GOE(FS)モデルについては SB の技術点 (FS) をそれぞれコントロール変数としている。

まず、不連続点をグループ境界とした場合をみる。男子では、構成点において上位グループダミーが１％水準で有意に推定されている。それに対して、基礎点とGOEについては上位グループダミーが有意に推定されていない。同様の傾向は女子でもみられ、構成点において上位グループダミーが１％水準で有意に推定されているのに対し、基礎点とGOEでは有意に推定されていない。

コントロール変数については、SP 総得点については男女とも全ての推定式において５％水準で有意に、シーズンベストについては、男子の基礎点とGOEを除いては10％水準で有意に推定されており、それぞれの得点が高いと FS の得点も高くなるという予想通りの結果となった[14]。

一方で、不連続点をグループ内に想定したモデル（表４中央）は、いわゆるプラシーボテストである。本研究における RDD 推定の目的は、グループの境界となる SP の順位で不連続点を確認することだが、逆に言えばグループの内部では、不連続点は生じないはずである。不連続点をグループ内部、すなわち SP 順位の５位と６位、７位と８位、11位と12位、13位と14位、17位と18位、19位と20位との間に設定し、RDD推定を行った結果、男女共に全ての推定式で有意な差はみられなかった。従って、能力をコントロールした上での不連続点、すなわちSP順位の差によるFSの得点の差は、グループ間の構成点でのみ観察され、グループ内では観察されない。

以上の結果より、ピア効果・発奮効果が、構成点、基礎点、GOE の全てに影響すると仮定するならば、構成点のみにおいて不連続性が発生する結果を説明できないため、構成点にのみ期待バイアスが存在すると解釈できる。すなわち、能力とは関係なく１つ上のグループに属しているだけで、構成点の評価が高まる点を示しており、平均的に男子で2.001点、女子で1.912点分、それぞれ高くなる。

Findlay and Ste-Marie (2004)においては、本研究の結果と異なり、技術点においても有意に期待バイアスが確認されている。これは、Findlay and Ste-Marie (2004)が、技術点においても相対評価を用いていた旧システム下における評価を対象としたためと考えられる。本分析で技術点に期待バイアスが確認されなかった点からは、評価システムの変更が少なくとも技術点において、主観的評価に内在するバイアスの軽減に寄与したと考えられる。

[14] シーズンベストについては、構成点、技術点それぞれをコントロール変数とした分析のほかに、FSのシーズンベストの合計点を変数に採用した分析も行ったが、推定結果に大きな違いはなかった。

6.2 ピア効果・発奮効果の検証

RDD による推定結果より、ピア効果・発奮効果が、構成点、基礎点、GOE の全てに影響すると仮定するならば、構成点のみで不連続性が発生する結果は説明できない。しかし、構成点のみで発生するピア効果・発奮効果の存在を、厳密に否定することは難しい。そこで、構成点について追加的な分析を行った。正のピア効果がある場合、実力が上の選手と同グループに入るとピア効果により評価の高い演技を行い、実力が下の選手と同じグループに入るとより評価の低い演技を行う(負のピア効果の場合は逆になる)と考えられる。すなわち、自分と他の選手の実力差に応じてピア効果の大きさが決まると考えられるため、ピア効果を「グループ内の他選手のシーズンベスト平均ー自己シーズンベスト」と定義して、(4)式に加えて分析を行った。基礎統計量(表３)においては、ピア効果の平均値の差が有意である。ピア効果の変数は、自分以外の選手のシーズンベストの平均値から自分のシーズンベストの値を差し引いて定義されるため、上位グループとの間においては値が大きく、下位グループとの間においては値が小さくなるためである。この変数は、発奮効果についての情報も有する。すなわち、グループ境界に近い選手が上位グループに入った時、上位進出への意欲が高まって評価の高い演技を行う一方、下位グループに入った時には上位進出へのモチベーションが下がって、評価の高くない演技を行う可能性がある。よって、発奮効果が存在する時、この変数の係数は正に推定されうる。

表４の右側は、構成点におけるピア効果を含めたモデルの推定結果である。これを、ピア効果を含めない推定（表４左）と比較すると、ピア効果を入れることで、男子では上位グループの係数が 2.001 から 2.332 と上昇し、有意水準は１％のままであった。また、女子では 1.912 から 1.730 に低下しているものの、10%水準で有意に推定された。一方、ピア効果の係数は有意ではない。その点から、不連続点を引き起こした主な要因が、ピア効果ではないと考えられる。従って、構成点におけるピア効果・発奮効果の影響は小さいと結論付けられる。

6.3 バイアスの影響についてのシミュレーション

次に、期待バイアスが実際の競技に与えた影響を確認するため、推定されたモデルを用いて、過去の大会の結果についてバイアスを除いた場合のシミュレーションを行い、順位の変動を検証した。バイアスを除いた推計総得点は、実際の総得点から、構成点で推定された不連続性を示す上位グループの係数を引いて求めた。具体的には、該当選手

表５　期待バイアスを取り除いた採点結果

世界選手権2010女子

選手名	グループ	実現値	実際順位	推計総得点	推計順位
浅田　真央	4	197.58	1	191.84	1
キム･ヨナ	3	190.79	2	186.97	2
安藤　美姫	**3**	**177.82**	**4**	**174.00**	**3**
C.ファヌフ	3	177.54	5	173.72	4
L.レピスト	4	178.62	3	172.88	5
C.コストナー	4	177.31	6	171.57	6

世界選手権2015男子

選手名	グループ	実現値	実際順位	推計総得点	推計順位
フローラン･アモディオ	3	229.62	9	225.62	9
閻　涵	4	229.15	10	223.15	10
小塚　崇彦	**1**	**222.69**	**12**	**222.69**	**11**
ジョシュア･ファリス	2	223.04	11	221.04	12
セルゲイ･ボロノフ	4	218.41	13	212.41	13
ロナルド･ラム	2	214.36	14	212.36	14

がグループ２に属していれば係数（男子 2.001 点、女子 1.912 点）を、グループ３であれば係数を２倍したものを、グループ４であれば係数を３倍したものを、それぞれ引いて求めた。その結果によるとバイアスの影響は大きく、14 大会・男女 28 競技中の 27 競技において順位の変動が生じ、上位８位(オリンピックにおける入賞圏内)に限っても、19 競技で発生する。以下、影響が大きかった２大会を例示する。まず、10 年世界選手権の女子では（表５上)、安藤美姫氏は FS の滑走順が第３グループで、実際の FS スコアは 177.82 点で最終順位４位であった。しかし期待バイアスを除去した場合の推計総得点は 174.00 点で３位となり、メダル圏内の成績であった。15 年世界選手権男子（表５下）の小塚崇彦氏は SP で出遅れ、第１グループでの演技となり、実際の成績は 222.69 点で第 12 位だった。だが、期待バイアスを除くと 11 位となる。16 年の世界選手権の日本人選手の出場枠が本大会の成績によって決定されたため、この順位の違いが日本の出場枠を変えた。その条件は、上位２人の順位の合計が 13 以下であれば３名、14－28 以下であれば２名である。日本人選手の首位は羽生結弦氏の２位、次位が小塚氏だったため、２名の合計は 14 で 16 年大会の出場枠は２となった。もし、期待バイアスが存在しなければ合計 13 となり出場枠は３であった。

16 年２月末時点における男子の世界ランキングでは、トップ 10 のうち３名が日本人であり、期待バイアスによる１枠の減少はメダル争いに大きな影響を及ぼした可能性がある。また、メダル争い以外の点においても、選手個人に経験を積ませる機会を失った

点は、選手強化という長期的視点から日本にとって大きな損失と考えられるため、主観的バイアスの存在がナショナルチームの育成戦略にも影響した可能性がある。

7. 結論と今後の課題

本研究は評価指標の主観性に焦点を当て、成果評価で発生するバイアスについて、フィギュアスケートの採点データによる検証を行った。その結果、審判の演技グループへの期待によって形成されるバイアスの存在が、構成点において確認された。一方、評価基準がより客観的な技術点の基礎点およびGOEでは同様の傾向はみられなかった。また、構成点について、グループ内で演技者のパフォーマンスが他の演技者に影響するようなピア効果や、上位グループに入ったことでモチベーションが高まる発奮効果を考慮した分析を行ったところ、同様の効果は確認されなかった。以上の分析結果は、より主観的な基準に依存する評価行動において、バイアスが生じる可能性が高いことを示す。さらに、期待バイアスの存在を考慮したシミュレーション結果によると、採点に対する期待バイアスの影響は大きく、場合によっては最終順位に影響を与えた可能性がある。

本分析結果は、主観的基準で発生しやすいシステマティックなエラー、すなわちバイアスが無視できない問題であることを示す。こうしたバイアスの存在は、労働者の納得感を低下させるなどの問題を引き起こす。日本でも成果主義的賃金体系が普及しつつある状況下において、正確性の高い人事評価のためには、主観的評価におけるバイアスの影響をなるべく小さくする必要がある。

主観的評価におけるバイアスの影響を軽減する方法としては、フィギュアスケートのように客観的評価と主観的評価の両方が利用可能な場合、客観的評価のウエイトを高くすることが考えられる。ISUは前述したオリンピックにおける審判スキャンダルを受け、04/05年シーズンより、技術点・芸術点のいずれについても、相対的な順位点から絶対的な評価点へと採点方式の大幅な変更を行った。ルール改正前のデータによる、Findlay and Ste-Marie (2004)の分析では、技術点・芸術点の両方に期待バイアスの存在が示唆されている。一方で、改正後のデータによる本研究の分析では、技術点における期待バイアスは確認できない。採点システムの変更により、技術点には客観性が増してバイアスが軽減され、競技全体としては評価がより正確になったと考えられる。

ただし、客観的な指標への過度なシフトは、マルチタスク問題を生み出す。労働者が客観的な評価が容易な業務に努力を傾倒させ過ぎると、最適な労働資源配分からの乖離が生じ、結果として企業の生産性が低下する。フィギュアスケートにおいては、選手が

技術点の向上に注力し過ぎて芸術面が軽視され、競技自体の魅力が低下しうる。

また評価バイアスを軽減するため、評価の透明性を高めること、すなわち評価の過程・結果の情報を公開し、人々の納得感を高めることも対策の１つとなりうるかもしれない。フィギュアスケートの国際大会においては、15/16 年シーズンまで審査に加わる審判の名前は公開されていても、各審判の評価は公表されていなかった 。これは、前述したスキャンダルを受け、審判を買収などの圧力から保護することを目的として 04/05 シーズンから導入された。Zitzewitz(2014)によると、この採点制度の匿名性が国籍バイアスを大きくした。また、Parsons et. al (2011)は、米国大リーグにおける審判の人種差別的な判定が、ビデオ判定システムの導入や観客動員が多さといった監視機能の高いケース、すなわち情報がより広く公開されているケースでは少なくなることを示した。これらの結果は、主観的な評価であっても評価過程の情報公開によって、バイアスが軽減する可能性を示唆する[15]。

実際、Shellenbarger (2016)によれば、米国では給与の不公平感を排除して、生産性を高めることを目的として、全従業員の給与に関する情報を社内で公開する雇用主が増えつつある。NTT コムリサーチ・日本経済新聞 (2015) による調査でも、人事評価に対する従業員へのフィードバックがある場合に、人事評価に対する満足度が高い傾向がある。一方で、給与公開や評価の過程を公開することは、過度の競争を組織内に引き起こす可能性や、評価や給与額が自己評価に合わない人々のモチベーションを引き下げ、生産性を下落させる可能性もあるため、こうした情報公開の動きによる評価におけるバイアスへの直接的な影響については、さらなる実証研究が必要である。

本研究で明らかになったように、主観的評価については期待や思い込みといった要素が強く作用しバイアスが発生しやすい。しかし、成果主義的賃金制度の導入をめぐる近年の議論では、主観的な評価は必ずしも悪いものであるとは考えられていない。労働者の業務遂行における複数の行動プロセスを、第三者が客観的に評価するのは難しいため、行動プロセスを日常的に把握している上司・同僚の主観的評価を活用することが求められている（高橋, 2004；阿部, 2006)。労働者の意欲および労働生産性の維持・向上させるため、客観的評価と主観的評価のバランス、主観的評価のバイアスの軽減によって評価に対する納得感を得るため、各企業・団体の特性を踏まえた制度設計が必要である。

最後に、今後の課題について述べる。まず、ピア効果・発奮効果と期待バイアスの識

[15] Zitzewitz (2014)やParsons et. al (2011)の分析は、本研究で明らかにした期待バイアスについての検証ではなく、認知バイアスの中でも内集団バイアスについての研究成果であることに注意されたい。一方で、透明性の向上を通じて評価者の緊張感を高めることは、期待バイアスの解消にも役立つ可能性も考えられる。

別の問題である。野口（2014）によれば、高い構成点を得るためには高い技術が必要であり、構成点に含まれる要素と技術点に含まれる要素には高い相関が存在する。従って、両効果が存在するならば、構成点、基礎点、GOE の全てにおいて観察されると考えられる。本研究では、構成点にのみ不連続点が確認されたこと、ならびに発生する両効果を明示的に導入したモデルでその影響が確認できなかったことから、ピア効果・発奮効果の影響は極めて小さいと考えている。しかし、本研究で定義したピア効果・発奮効果変数は両効果を完全に表していない可能性があり、その意味において、構成点にのみ発生するピア効果・発奮効果の存在を完全に否定していない。技術点・構成点の相関についての詳細な検証や、ピア効果・発奮効果が発生する状況を特定した分析が求められる。

また、審判の経験が期待バイアスに大きな影響を及ぼすとき、その発生メカニズムを検証することも重要である。すなわち、経験の多い審判ほど自分の経験に依存して判定するためにバイアスが大きくなるのか、経験の浅い審判ほど正確な判定が行えずにバイアスが大きくなるのかといった検証は、評価システムを設計する上でも有用であろう。

参考文献

阿部正浩（2006）「成果主義導入の背景とその功罪」『日本労働研究雑誌』No. 554, pp. 18-37.

荒川静香（2006）「金メダルの秘密」『Sports Graphic Number』No.649(2006 年 3 月 16 日).

荒川静香（2013）『誰も語らなかった　知って感じるフィギュアスケート観戦術』朝日新書.

NTT コムリサーチ・日本経済新聞（2015）『人事評価に関する調査』
<http://research.nttcoms.com/database/data/001961>（2017 年 11 月 30 日閲覧).

大竹文雄・唐渡広志（2003）「成果主義的賃金制度と労働意欲」『経済研究』Vol.54, No.3, pp. 193-205.

大湾秀雄（2011）「評価制度の経済学 : 設計上の問題を理解する」『日本労働研究雑誌』,Vol.53, No.12, pp. 6-21.

高橋伸夫（2004）『虚妄の成果主義 : 日本型年功制復活のススメ』日経 BP 社.

野口美惠（2014）「フィギュア採点の最難関を徹底解説！ソチ演技で見る『演技構成点』とは。」『Number Web』文藝春秋, <http://number.bunshun.jp/articles/-/802055>（2017 年 11 月 30 日閲覧).

樋口美雄（2006)「人事経済学からみた成果主義人事の制度設計とその運用」樋口美雄・八代尚宏・日本経済研究センター編著『人事経済学と成果主義』日本評論社, pp.11-26.

森津太子（2008）「検索容易性の経験が社会・認知的判断に及ぼす効果」『放送大学研究年報』No.26, pp.47-54.

Angrist, J. D. and J. S. Pischke (2009) *Mostly Harmless Econometrics*, Princeton University Press.

Bruine de Bruin, W. J. A. (2005) "Save the Last Dance for Me: Unwanted Serial Position Effects in Jury Evaluations," *Acta Psychologica*, Vol.118, pp.245-260.

Campbell, B. and J. W. Galbraith (1996) "Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments," *Journal of the Royal Statistical Society*, Vol.45, No.4, pp.521-526.

Deci, E. L. (1975) *Intrinsic Motivation*, Plenum Press, New York.

Elvira, M. and T. Robert (2001) "The Effects of Race and Worker Productivity on Performance Evaluation," *Industrial Relations*, Vol.40, No.4, pp.571-590.

Findlay, L. C. and D. M. Ste-Marie (2004) "A Reputation Bias in Skating Judging, *"Journal of Sport and Exercise Psychology*, Vol.26, pp.154-166.

Ikegami, T. and G. Ganesh (2014) "Watching Novice Action Degrades Expert Motor Performance: Causation between Action Production and Outcome Prediction of Observed Actions by Humans," *Scientific Reports*, Vol.4.

Imbens, G. W. and J. M. Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, Vol.47, No.1, pp.5-86.

International Skating Union (2014) "Special Regulations & Technical Rules Single & Pair Skating & Ice Dance 2014"
<http://static.isu.org/media/165218/2014-special-regulation-sandp-and-ice-dance-and-technical-rules-sandp-and-id.pdf>（2017年7月14日閲覧）.

Kahneman, D. (2012) *Thinking, Fast and Slow*, Penguin Books.

Morgan, H. N. and K. W. Rotthoff (2014) "The Harder the Task, the Higher the Score: Findings of a Difficulty Bias," *Economic Inquiry*, Vol. 52, No.3, pp.1014-1026.

Ofir, C., P. Raghubir, K. B. Monroe and A. Heiman (2008) "Memory-Based Store Price Judgments: The Role of Knowledge and Shopping Experience," *Journal of Retailing*, Vol.84, No.4, pp.414-423.

Parsons, C. A., J. Sulaeman, M. C. Yates and D. S. Hamermesh (2011) "Strike Three: Discrimination, Incentives and Evaluation," *American Economic Review*, Vol.101, No.4, pp.1410-1435.

Price, J. and J. Wolfers (2010) "Racial Discrimination among NBA Referees," *Quarterly*

Journal of Economics, Vol.125, No.4, pp.1859-87.

Shellenbarger, S. (2016) "Open Salaries: the Good, the Bad and the Awkward" *The Wall Street Journal*, 12 January.
<https://www.wsj.com/articles/open-salaries-the-good-the-bad-and-the-awkward-1452624480> (2017年11月30日閲覧).

Tversky, A. and D. Kahneman (1974) "Judgment under Uncertainty: Heuristics and Biases," *Science*, Vol.185, No.4157, pp.1124-1131.

Yamane, S. and R. Hayashi (2015) "Peer Effects among Swimmers," *The Scandinavian Journal of Economics*, Vol.117, No.4, pp.1230-1255.

Zitzewitz, E. (2014) "Does Transparency Reduce Favoritism and Corruption? Evidence from the Reform of Figure Skating Judging," *Journal of Sports Economics*, Vol.15, No.1, pp.3-30.