



RIETI Discussion Paper Series 26-J-010

順序は評価を歪めるのか： 大規模ピアノコンクールの観察データと フィールド実験からのエビデンス

浅川 慎介
佐賀大学

中室 牧子
経済産業研究所

山口 慎太郎
東京大学



Research Institute of Economy, Trade & Industry, IAA

独立行政法人経済産業研究所

<https://www.rieti.go.jp/jp/>

順序は評価を歪めるのか：**大規模ピアノコンクールの観察データとフィールド実験からのエビデンス¹**

浅川慎介（佐賀大学） 中室牧子（慶應義塾大学・RIETI） 山口慎太郎（東京大学）

要 旨

本研究は、専門家による逐次評価において、評価対象の提示順が判断を歪める「順序効果」がなぜ生じ、またそれを是正することが可能なかを検証する。日本最大規模のピアノコンクールを対象に、2004～2022 年度の長期観察データと、2023 年度に実施したクラスターランダム化比較試験（RCT）を組み合わせ分析を行った。観察データの分析から、演奏順が早い奏者ほど評価が低くなる頑健な順序効果が一貫して確認された。この順序効果は、参加者数が少ない大会や、演奏者の水準が高く競争の激しい上級クラスにおいてより顕著であった。これらの結果は、審査員が評価全体の整合性を保とうとする動機から、全体のパフォーマンス分布を把握できない序盤において極端な評価を避け、初期の評価を抑制する「キャリブレーション」によって順序効果が生じている可能性を示唆している。これを踏まえ、本研究では、審査員に対して過去データに基づく順序効果の存在と大きさを明示的に伝える情報提供介入を設計し、その因果効果を RCT によって検証した。分析の結果、全体としては情報提供によって順序効果が有意に是正されたという強い証拠は得られなかった。ただし、上級クラスに限定すると、序盤の不利が部分的に緩和される可能性が示された。これらの結果は、逐次評価における順序効果が頑健に存在する一方で、注意喚起や情報提供といった個人の努力に依存した介入のみでは十分に是正できず、制度上の対応が必要であることが示唆される。

キーワード：ランダム化比較試験、逐次評価、順序効果、カリブレーション

JEL classification: I0, I2

RIETI ディスカッション・ペーパーは、専門論文の形式でまとめられた研究成果を公開し、活発な議論を喚起することを目的としています。論文に述べられている見解は執筆者個人の責任で発表するものであり、所属する組織及び（独）経済産業研究所としての見解を示すものではありません。

¹本稿は独立行政法人経済産業研究所（RIETI）におけるプロジェクト「機能する EBPM の実現に向けた総合的研究」の成果の一部である。本稿の原案は RIETI のディスカッション・ペーパー検討会では、大竹文雄氏、川口大司氏他、検討会参加者からの有益なコメントに感謝したい。また、本稿の執筆にあたって、一般社団法人全日本ピアノ指導者協会（ビティナ）から多大な支援を頂いた。ここに記して、感謝の意を表したい。本研究は、SFC 研究倫理審査委員会の承認を受けて実施された（承認番号:485, 2023 年 4 月 9 承認）。また本研究にあたっては、東京財団政策研究所（研究課題：教育の「質」が子供の学力や非認知能力に与える影響）および慶應義塾大学学事振興資金からの支援を受けた。

Introduction

1人ないしは複数の専門家が、異なる個人が連続して行うパフォーマンスを評価する逐次評価 (Sequential Evaluation) は現実社会では頻繁に見られる設定である。例えば、企業の採用面接、音楽やスポーツのコンテスト、大学の口頭試験など多岐にわたる。このような逐次評価では、個人のパフォーマンスを独立して評価する場合とは異なる結果になることが多くの研究で示されている (Simonson, 1990; Read & Loewenstein, 1995; Stewart et al., 2006; Sunstein et al., 2001; Simonsohn & Gino, 2013; Bhargava & Fisman, 2014; Chen et al., 2016; Leibovitch, 2016; Hartzmark & Shue, 2018)。一連の研究の含意は、専門家の評価が「真のパフォーマンスの質」を反映しておらず、評価の順番が評価者の認知、疲労、動機づけなどによって、評価に系統的な偏りを生じさせる可能性を指摘している。つまり、同じパフォーマンスであったとしても、提示される順番によって異なる評価や判断になるというわけだ。逐次評価における順番が評価に与える影響は既に様々な分野で研究が行われており、過去の文献で順序効果 (Serial Position Effect) または逐次バイアス (Sequential Bias) と呼ばれている。本研究でも先行研究に倣って順序効果と呼称する¹。

順序効果に関する過去の研究には、ピアノやバイオリンなどの音楽コンクール (Flôres & Ginsburgh, 1996; Glejser & Heyndels, 2001; Bruine de Bruin, 2005a; 松山, 2021)、アイドル、ダンス、お笑いなどのテレビ放映されるコンテスト (Page & Page, 2010; Collins, et al. 2019; Arai & Okazawa, 2025)、フィギュアスケートやシンクロナイズドスイミングなどのスポーツ競技 (Bruine de Bruin, 2005b; 2006a) について調べた研究がある。より広く現実社会に即した設定として、採用面接 (Vives et al. 2021)、学術誌の査読 (Orazbayev, 2017)、飲食店の衛生検査 (Ibanez & Toffel, 2020)、受刑者の仮釈放 (Plonsky et al. 2023; Danziger et al. 2011)、ワインやウイスキーなど酒のテイスティング (Mantonakis et al. 2009; Quigley-McBride et al. 2018)、ビジネスのピッチや助成金の審査 (Clingsmith & Shane, 2017; Bian et al. 2022)、裁判での判決、ローンの審査、メジャーリーグの審判 (Chen et al. 2016) においても、観察データを用いた研究が行われ、順序効果は決して小さくないことが示されている。

本研究の目的は、日本で最大のピアノコンクールの演奏者と審査員の観察データを用いて順序効果が存在するかを検証し、それを緩和することができるかを実証的に検証することにある。まず、2004～2022 年度に実施された観察データを用いて順序効果の推定を行い、その背後にあるメカニズムの検討も行った。更に、2023 年度に全国で

¹ Asmat et al. (2023) は、経済学では通常「バイアス」とは、客観的な能力差では正当化できない特定の集団への偏った扱いを意味するが、この分野の文献では、「バイアス」と「嗜好」を明確に区別することは難しいことを指摘している。

開催されたコンクールをフィールドとして、大会を単位とするクラスターランダム化比較試験を実施し、順序効果に関する情報提供が審査員の評価に与える因果効果を推定した。

本研究の結果、明らかになったことは次の通りである。第一に、観察データの分析からは頑健な順序効果が確認され、特に前の方の演奏者が不利になっていることが示された。1 番最初の演奏者の評価が最も低く、11 番目以降の演奏者と比較すると $0.117\ s.d.$ も評価得点が低くなっている。この傾向は、参加者人数が少ない大会や、上級者が参加する大会で大きい。

第二に、前の演奏者が不利になるメカニズムとして、単なる疲労の副作用ではなく、審査員が評価全体の一貫性を保ちたい、あるいは矛盾を避けたいという動機によって生じている可能性が示唆された。つまり、パフォーマンスの質の全体の分布を把握できない初期段階に、極端な評価を避け、後から来る演奏者との整合性を保とうとして評価を調整するキャリブレーション (Calibration) と呼ばれる行動がみられている。しかし、審査員の経験豊富さやリハーサル開催によって、キャリブレーションが促進され順序効果が緩和されるという証拠は得られなかった。

第三に、順序効果を緩和するために審査員に情報提供し、その因果効果を計測したランダム化比較試験の結果においても、介入が順序効果を緩和したという強い証拠は得られなかった。しかし、上級の演奏者を中心とした大会では、介入によって順序効果の約半分を緩和したことが確認された。本人に対する注意喚起や努力に頼って順序効果を緩和するのは難しいことが明らかとなった。

本研究は、既存研究に対して以下の点で貢献する。第一に、これまでの研究の多くが小規模かつ短期間のデータを用いてきたのに対し、本研究は日本最大規模のピアノコンクールにおける大規模かつ長期の観察データを用いて、逐次評価における順序効果が一時的な現象ではなく、制度的に頑健に生じていることを示した点である。第二に、順序効果が「後ろが有利になる」疲労型のパターンではなく、「前の評価が抑制される」という形で現れることを明らかにし、そのメカニズムとしてキャリブレーションが重要な役割を果たしている可能性を示した点である。第三に、順序効果の緩和については理論的または仮想的な実験に基づく検証しか行われていない中、実際のコンクールをフィールドとしてクラスターランダム化比較試験を実施し、情報提供介入の因果効果とその限界を実証的に検証した点である。

本論文の構成は以下の通りである。第 2 章では文献のレビュー、第 3 章では本研究で用いるピアノコンクールおよびそこから得られた匿名加工データの概要を説明する、第 4 章では観察データを用いた分析について、第 5 章ではフィールド実験の概要とその分析の結果を述べる。第 6 章は結論である。

Literature Review

逐次評価が行われる状況では、評価対象の質が同一であっても、提示される順序によって評価結果が体系的に変化する「順序効果」が生じることが、これまでの研究によって広く確認されている。こうした順序効果は、フィギュアスケートやシンクロナイズドスイミングなどのスポーツ競技 (Bruine de Bruin, 2005b; 2006a; 2006b; Rotthoff, 2015)、音楽コンクール (Flôres & Ginsburgh, 1996; Glejser & Heyndels, 2001; Bruine de Bruin, 2005a; 松山, 2021)、採用や昇進の判断 (Vives et al. 2021)、学術誌の査読 (Orazbayev, 2017)、飲食店の衛生検査 (Ibanez & Toffel, 2020)、受刑者の仮釈放 (Plonsky et al. 2023; Danziger et al. 2011)、ワインやウイスキーなど酒のテイスティング (Mantonakis et al. 2009; Quigley-McBride et al. 2018)、ビジネスのピッチや助成金の審査 (Clingsmith & Shane, 2017; Bian et al. 2022)、裁判での判決、ローンの審査、メジャーリーグの審判 (Chen et al. 2016) など、多様な分野において観察されており、評価が逐次的に行われる限り、専門家による判断であっても例外ではないことが示されてきた。逐次的な評価構造そのものが、判断や結果を歪める可能性が示唆されている。

過去の研究は、順番が評価に影響を与えるということについては一致しているが、評価の方法や状況によって、どの順番であれば有利になるのかになるのかは一致を見えていない。逐次評価における順序効果を理論的に説明した最も影響力のある枠組みの一つが、Hogarth and Einhorn (1992) による信念調整モデル (Belief-Adjustment Model) である。これは、人が逐次的に情報を受け取る際、現在の信念をアンカーとして新たな情報を調整的に統合するという認知過程に基づき、なぜ前の方が有利になったり (Primacy Effect) や後ろの方が有利になったり (Recency Effect) が生じるのかを統一的に説明し、以降の順序効果に関する研究の基礎をなしている。その後続いた多くの実証研究は、順番が「後ろのほうが有利」という結果を示しているものが多い一方 (Bruine de Bruin, 2005b; 2006a; 2006b; Rotthoff, 2015; Collins et al. 2019; Glejser & Heyndels, 2001; 松山, 2021; Bruine de Bruin, 2005a; Page & Page, 2010; Quigley-McBride et al. 2018; Plonsky et al. 2023)、後ろのほうが有利なのではなく、「前のほうが不利」であることを示した研究もある (Clingsmith & Shane, 2017; Bian et al. 2022)。逆に、初めのほうが有利であることを示した研究もある (Mantonakis et al. 2009; Arai & Okazawa, 2025)。

本研究と同様に、ピアノコンクールを対象にした研究は、全て世界3大コンクールの1つと言われるエリザベート王妃国際音楽コンクールのデータを用いている (Flôres & Ginsburgh, 1996; Glejser & Heyndels, 2001; Ginsburgh & van Ours,

2003)。このコンクールは、本選に12名が出場し、1日に2名が演奏、計6日間で行われる。3つの研究はすべて1950年代から1990年代にかけての11または12回のコンクールのデータを用いて分析している。3つの研究では一貫して、初日が不利になるという順序効果があることが確認されている。また同日に演奏した場合でも、1人目が不利になるという。しかも、Ginsburgh & van Ours (2003)によれば、音楽コンクールにおける順位が、その場限りの評価ではなく、後続の演奏機会や収益を通じて音楽家としての長期的な成功に影響を与えることを明らかにしている。

ただし、これらの研究には、限界もある。エリザベート王妃国際音楽コンクールは、個々の審査員の点数が非公開のため、審査員の点数を合算した後の結果しか観察することができず、評価が逐次的に行われる過程そのものがどのように判断を歪めるのかというメカニズムを直接検証していない。加えて、演奏者のサンプルサイズが約140名と小さく、演奏順が真に外生的に決まっているかどうかという点については十分な検討が行われていない。

どうして順序効果が生じるのか。先行研究では、順序効果が生じる原因は、大きく分けると「疲労」と「動機」とであると指摘されてきた (Plonsky et al. 2023)。

疲労は様々な形で判断や評価に影響を与える。疲労による順序効果は、受刑者の保釈に対する判断 (Danziger et al., 2011)、飲食店の衛生検査 (Ibanez & Toffel, 2020)、論文の査読 (Orazbayev, 2017) など様々な設定で確認されている。いずれも評価者が肉体的に消耗した結果、よりデフォルトに近い判断や評価をしやすくなることによって生じると解釈されている。しかし、疲労仮説で順序効果のすべてを説明できるわけではない。審査の間に休憩を挟んで疲労を軽減したとしても、順序効果は残ると指摘する研究もあるほか (Glejser & Heyndels, 2001; Plonsky et al. 2023)、時間が経つほどデフォルトとは逆の判断をしていることを示す研究もあり、疲労では説明できない結果となっている研究も多い (Plonsky et al. 2023)。

もう1つの可能性は、評価者が逐次評価において「内的一貫性 (internal consistency) を保ちたい」または「評価の矛盾を避けたい」という動機によって生じているというものである²。Plonsky et al. (2023)は、こうした動機付けの根底には、(1)戦略的一貫性 (Strategic Consistency)、(2) 割り当て (Quota)、(3) キャリブレーション (Calibration) という3つのメカニズムが存在するという。

²評価者の一貫性を保ちたいという内的動機を扱った経済学の論文に、Falk & Zimmermann (2018)がある。この研究は、人々が最初に表明した判断や意見に「コミット」すること自体が、その後の情報処理を歪めることを、実験室実験によって示した。著者らは、人は一度表明した判断と矛盾する行動をとることに心理的コストを感じるため、新しい情報を過小評価し、最初の判断に引き寄せられたまま意思決定を行ってしまうからだと解釈している。追加実験により、この効果は他者からの視線がなくても生じることが示されていることから、主に内的動機によって生じることが確認されている。

(1)の戦略的一貫性は、評価者の判断が、例えば再審、控訴、不服申立などによって、のちに第三者から再検証 (scrutiny) される可能性がある場合、評価者がのちに再検証の対象となりやすい厳しい判断 (却下・有罪など) を避けたいという動機から生じる。逐次評価の順番がすすむほど、過去の判断との整合性が問われやすくなるため、順番が後ろになるほど再検証されにくい寛大な判断を選びやすくなると予測される。

(2)の割り当ては、逐次評価の中で例えば合否のような判断をする場合、合格者の割合について、例えば「合格者は全体の 10%まで」のように、何らかの割り当てがある場合が想定される。もし仮にある評価者が逐次評価の最初の方で多くの合格を出してしまったら、後ろの方では合格という判断をしにくくなってしまうというようなことである。順番が後ろになればなるほど枠が少なくなることによって厳しい評価にならざるを得ないことから、序盤よりも後半の方が不利になると予測される。

割り当ては、Chen et al. (2016)が指摘する「ギャンブラーの誤謬」(Gambler's Fallacy) と区別をつけにくい。ギャンブラーの誤謬とは、本来は独立に生じる事象について、短期的には結果が均されるはずだ (Mean Reversion) と誤って信じてしまう認知的バイアスのことを指す。例えば直前に合格が数回続くと、次は不合格だろうと考えてしまい、直前の判断と反対の決定を下しやすくなる。観測される結果だけを見ると、割り当てとギャンブラーの誤謬はいずれも類似した順序効果を生じさせ、両者を完全に識別することは困難である。しかし、Chen et al. (2016)は、「直近の判断が過度に強い影響を持つかどうか」を検証することで、両者を区別することを試みている。

(3)のキャリブレーションは、評価者が逐次的に判断を行う中で、自分自身の評価基準をこれまでに観察したケースに合わせて調整していくプロセスを指す (Plonsky et al. 2023)。つまり、評価者が最初から固定された基準を持っているわけではなく、順番が進むにつれて評価基準が内生的に更新されていく。評価者は、逐次評価の初期段階では全体のパフォーマンスの質の分布がどうなっているかをまだ十分に把握していないため、序盤では例えば最低評価や最高評価などの極端な判断や評価を避ける。そして、後半になって全体の分布が見えてきて、判断基準が固まってくると初めて最低評価や最高評価を含む大胆な判断をするようになるというわけだ (Unkelbach & Memmert, 2014)。キャリブレーションはコンテストの審査という文脈でよく見られることが知られている (Antipov & Pokryshevskaya, 2017; Bian et al. 2022; Bruin de Bruin, 2005b; Rotthoff, 2015)。

こうした順序効果が観察される原因は評価者側ではなく、被評価者側にある可能性もある。例えば、最初に評価の対象となることで不安や緊張を感じ、十分なパフォーマンスを発揮できないということも考えられる。しかし、過去の研究は、評価の対象

が文書資料や論文であり、被評価者が評価される順番を知らないケースにおいても順序効果が生じていることを示しており (Bian et al. 2022; Orazbayev, 2017)、順序効果の原因は評価者側の疲労や動機にあるとする文献が多い。

以上のように、逐次評価における順序効果の存在自体は多くの研究によって示されてきたものの、そのメカニズムについては疲労や動機など複数の仮説が併存しており、特に専門家が評価する高ステークスな競争においてどの要因が支配的であるかは十分に明らかにされていない。

評価者に対する介入によって、順序効果を緩和することができるかという検討は少ないながら、既に始まっている。第 1 に、経験のある評価者を起用するということだ。経験豊富なマネージャーほどこれまでの行動、特に行動が明示的かつ公に行われる場合、より一貫性がある傾向があることが示唆されている (Finkelstein & Hambrick, 1990)。加えて、専門知識の程度が高いほど、タスクのばらつきが小さくなることが示唆されている (Haerem & Rau, 2007)。もし順序効果がキャリブレーションによって生じているならば、評価者側が全体のパフォーマンスの質の分布について何らかの情報を持っている場合、順序効果は生じにくいかもしれない。ところが、経験豊富な審査員が順序効果を緩和できるのかについては一致を見ていない。Bian et al. (2022)や Chen et al. (2016)は、過年度の審査員の経験が長くなるほど、順序効果が小さくなることを明らかにしている。一方、Plonsky et al. (2023)や Unkelbach et al. (2012)は、より経験豊富な審査員ほど評価の内的一貫性を重視する傾向があるため、かえって順序効果は大きくなることを示している。

第 2 に、Akl & Tewfik (2014, 2016)の理論的なアプローチがある。認知バイアスの影響を組み込んだ人間の意思決定の数学的モデルを開発し、評価者の認知バイアスを最小限に抑えるように情報の提示順序を最適化することが検討された (2014)。その上で、およびバイアスを軽減する介入設計 (2016) として、過去に観測されたデータに基づいて、提示する情報そのものを修正する方法や、特定の観測順序に対して、人間側に判断境界 (閾値) を修正するよう訓練する方法などを提示し、「順序を変えるだけで判断の精度が改善する」という実証的含意を与えた。Wang & Pananjady (2022)もまた逐次評価における順序効果をモデル化し、観測された評価スコアから真のランキングを推定し、逐次バイアスを修正するための効率的なアルゴリズムを提案している。クラウドソーシングデータを用いた検証では、このアルゴリズムによって順序バイアスを修正したランキングが、真のランキングとより正確に一致することを示した。しかし、Akl & Tewfik (2014; 2016)や Wang & Pananjady (2022)はいずれも理論的な検討や仮想的な実験に止まっており、現実社会における高ステークスな意思決定において、順序効果を緩和しうるかを明らかにするための検証はほとんど存在しな

い。本研究は、このギャップを埋めることを目的として、日本最大規模のピアノコンクールの大規模な観察データとフィールド実験を組み合わせ、逐次評価における順序効果の存在を確認し、その緩和の可能性について探る。

Institutional Background

本研究では、一般社団法人「一般社団法人全日本ピアノ指導者協会」（以降、ピティナ）から提供された匿名加工データを用いる。ピティナには、約 14,000 人の指導者が登録しており、その指導者の下で約 350,000 人の生徒が学んでいる。日本最大規模のピアノコンクールであるピティナ・ピアノコンペティションを運営している団体である。ピティナは毎年 4 月 1 日から募集が始まり、5～7 月の 2 か月間に地区予選、7～8 月に地区本選、8 月に全国大会が行われる。全国各地で 300 か所以上の地域でコンクールを実施する。各地域で行われるコンクールは「大会」と呼ばれる。

各大会に参加する演奏者は、「級」に分かれて演奏を行う。級は学齢で区分されており、就学前児童の A2 級、小学 2 年生以下の A1 級、小学 4 年生以下の B 級、小学 6 年生以下の C 級、中学 2 年生以下の D 級、高校 1 年生以下の E 級、高校 3 年生以下の F 級、15 歳以下の Jr. G 級、22 歳以下の G 級、年齢制限なしの Pre 特級、特級がある。上級になると、演奏時間も長く、大規模な国際ピアノコンクールにおいて上位に入賞した経験のある演奏者も参加し、国際コンクール出場のための登竜門とみなされる高ステークスな競争となる。1 人につき、2 大会まで参加可能であるが、ソロ部門では異なる級の併願はできない。各大会・級の参加者数の分布は図 1 で示された通りである。

地区予選、地区本選、全国大会の 3 つからなるトーナメント形式である。ソロ、デュオ、グランミューズと 3 つの部門がある。年齢層の低い参加者が多い A2～F 級は、地区予選に合格した者が地区本選に進み、地区本選で合格した者が全国大会に進む。年齢制限のない Jr. G 級以上は、選抜段階が級によって異なるが、予選から本選に進むことは変わらない。開催時間は大会によって異なるが、低い級が午前～夕方、高い級が夕方～夜間に演奏することが多い。

予選、本選を勝ち抜いたのちに全国大会で上位に行けば、最上級の特級の 1 等が 150 万円と高額な賞金を得られる。Pre 特級と G 級の 1 等が 40 万円、Jr. G 級が 25 万円、F 級の 1 等が 20 万円、E 級の 1 等が 15 万円、D 級の 1 等が 10 万円、C 級以下は楯・賞状などが授与される。これ以外にも、B 級、C 級の優秀賞受賞者は一部の大学の入学金が免除される。

各大会において、演奏者はあらかじめ決められた順番で一人ずつ演奏し、審査員は各演奏が終了するごとに評価を行うが、過去の評価を修正することは可能である。審

査員は十分な指導・演奏経験を積んだピアノ指導者・プロのピアニストで、原則として大会が行われる地域に在住する指導者ではない音楽家が務める。1つの大会・級の審査に、5名の審査員が割り当てられる。この5名の組み合わせも、コンピュータによってランダムに決定されるが、一部には審査員の都合によって調整が行われることもある。同じ5名の審査員が、同じ大会を最初から最後まで通して採点を行う。各審査員が10点満点で、0.1点刻みで採点する。審査員には十分な専門的裁量が与えられており、評価基準は明示的な数式やアルゴリズムによって機械的に定められているわけではない。各演奏者の評点は、審査員の評点の平均点として算出される。その上位を次選進出者（合格）とする。次選に進出できる演奏者の数は各大会・級で予め定められている。例えば、予選のA2～F級の場合、上位35～45%を次選に出場させるという明示的な割り当てがある。また、審査員は、本選に出場を推薦できる閾値はあらかじめ伝えられている。審査基準は極めて厳格に決定されており、すべて公表されている。地区予選における各演奏者の評点の分布は Appendix 1 に示されるとおりで、年によるばらつきは見られない。

予選の審査については、参加者は演奏番号のみで審査され、審査員に対して氏名や年齢などの情報は一切知らされることはない。審査員は他の審査員の審査結果を見ることはできず、全員がそれぞれに独立した判断をする。昼食や休憩中に評価について話し合うことは審査上のルールによって明確に禁じられているため、審査員間のピア効果や社会的プレッシャーの影響が生じることは考えにくい。また、演奏会場には、必ず運営管理者が立ち会い、審査や演奏の様子をモニタリングしていたため、会場によってルールや運用が異なるということは生じない。

Data

本研究では、ソロの予選のみをメインの分析の対象とする³。表1は、2004～2022年度（コロナウィルス感染症対策で中止となった2020年度を除く）の参加者および審査員の属性の記述統計量である。82.9%の演奏者と80.1%の指導者が共に女性である。エントリー回数の平均は1.23回となっており、同年に開催された大会・級で1回以上エントリーしている演奏者もいる。指導者の平均経験年数は18.9年で、経験のある指導者が多い。

このコンクールは日本最大規模であることから、サンプルサイズが大きい。審査員・奏者にユニークIDが付番されており、年齢や性別などの属性についての情報も含

³ ソロ予選は、すべての演奏者が同一形式・同一評価基準の下で評価され、デュオやグランミューズ部門に比べて演奏順以外の制度的な異質性が小さいため、逐次評価における順序効果を識別するのに最も適している。

まれていることから、逐次評価における順序効果の推定に当たって極めて有用なデータである。これに加えて、このデータを用いる最も大きな利点は、演奏順がコンピュータによってランダムに決定されており、演奏者は大会当日に会場で演奏順が知らされ、演奏順を変更することはできないということである。この制度設計と後述するバランステストの結果を踏まえると、演奏順は演奏者や審査員の観察可能な属性とは独立に決定していると考えられる。一方、過去の研究には、パフォーマンスの順番がランダムに決定していることが保証されていないケースも多い。特に、テレビ放映されるようなイベントでは、視聴率を高めるために人気のあるパフォーマーを最初に出場させたり、あるいはクライマックスである最後に出場させるインセンティブがある (Page & Page, 2010; Collins, et al. 2019)。しかし、このコンクールではそうした可能性は排除されており、演奏順の因果効果を識別する上で有用である。

表 3 は、演奏順が演奏者および審査員の属性と相関していないことを確認するためのバランステストの結果を示している。いずれの属性についても統計的に有意な相関は確認されず、演奏順は実質的にランダムに決定されていると考えられる。特に、審査員の属性も演奏順と相関がないことは重要である。このことは、順序効果が審査員の属性によって生じる偏見や差別（例：審査員と同じ性別の演奏者を高く評価する、など）によるものである可能性は低いことを意味する。

Main Results

順序効果は生じているのか

まず、2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までの 17 年間にピティナ・ピアノコンペティションに参加したのべ約 39 万人約 185 万回の演奏（リハーサル除く）と 1,067 名の審査員の観察データを用いて、順序効果の存在を確認する。

評点は、5 人の審査員の平均点を、開催年・大会・級ごとに平均 0、分散 1 となるように標準化した（以降、評点と記載する場合は、この標準化得点を指すものとする）。これを演奏順が 1 番から順番に並べたのが図 2 である。これを見ると、最初の 10 人程度の演奏者の評点は統計的に有意に低く、その後は演奏順による体系的な差は確認されない。特に、最初の 3 番目までの評点が低いのが顕著であり、順序効果の大部分が序盤に集中している。つまりこのデータにおいては、Clingsmith & Shane (2017) や Bian et al. (2022) と同様に、「前のほうが不利になる」という順序効果が観察される。

このことをより厳密に確認するため、評点を被説明変数として、演奏者の大会・級別の 10 番目までの演奏順（欠席者・リハーサルを除く）のダミー変数で回帰したのが

表4である。なお、すべての推定において、演奏者、審査員、大会開催年・地区、級、累計エントリー回数（リハーサルを含む）の固定効果、リハーサルをのぞく大会・級別の演奏者数、リハーサルがある大会・級ダミーを制御した。これをみると、最初の演奏者において最も大きな不利が観察され、その影響は2番目、3番目と急速に小さくなる。4番目以降では隣接順位間の差は統計的に有意ではなく、順序効果は主として最初の1～3番目に集中していることがわかる。1列目のベースラインの推定をやや詳細に見てみると、最初の演奏者は、11番目以降の参照カテゴリーの演奏者と比較すると0.117 *s.d.*も評点が低く大きな順序効果が生じていることがわかる。演奏順の係数は、順番があとになればなるほど徐々に小さくなり、10番目になると1番目の20分の1程度の大きさとなり、統計的に有意ではない。また、最終演奏番号の係数も統計的に有意ではなく、最初の方の順番が不利であることは確認できるが、最後の方が有利になっているという証拠は得られない。

表1の8列目の素点を被説明変数にした推定においても、ほとんど同じパターンとなっている。本コンクールでは、5名の審査員が0.1点刻みで採点し、その平均点によって合否が決定される。このため、1人の審査員の評価が0.1点変わるだけで、平均点は0.02点変動する。例えば演奏者の平均点が予選通過の閾値よりも0.001点下の場合、0.02点の差は逆転をもたらすのだが、0.2点下の場合、0.02点では逆転しない。このため、演奏者が予選通過の閾値付近の評点であれば、0.02点の差は結果を左右しうる。このことを元に、表4の素点で推定された結果が0.02点よりも統計的に有意に大きいかどうかをWald検定でみてみると、1番目と2番目の演奏者の順序効果は0.02点よりも統計的に有意に大きい（1番目：*t*値=17.0、2番目：*t*値=5.0）、3番目以降は統計的に有意な差はない。

次に、どのようなケースで順序効果が大きくなるかを検討する。参加者数による異質性を見たのが、表4の2列目（10人以上）と3列目（9人以下）である。結果をみると、10人以上の大規模な大会・級における順序効果は表4の1列目のベースラインのそれとほとんど変わらない。一方、9人以下の小規模な大会・級では推定値の標準誤差は大きくなるものの、係数の大きさはほとんどすべての順番でベースラインの2倍以上となっており、7番目までは係数差も統計的に有意である。なお、Appendix 2では、大会・級参加者数が10人よりも多くなった場合に、結果に影響を与えるかを確認している。大会・級参加者数が70人を超えるような大規模な大会・級はそもそも数が少ないため、標準誤差が大きく、明確な順序効果は見られないが、総じて参加者数が少ない方が順序効果は大きい傾向があることがわかる。

級による異質性をみたのが表1の5列目（F級以下）と6列目（G级以上）である。既に述べた通り、G级以上の演奏者にはプロを目指すものも多く、参加者の年齢

もパフォーマンスの質も高い。結果をみると、F 級以下の順序効果はベースラインのそれとほとんど変わらない。しかし、G 級以上の演奏順の係数はかなり大きい。その差も有意な 4 番目までと 6 番目には統計的に有意である。G 級以上は国際コンクールへの登竜門とみなされており、審査は特に難易度が高く、評価者にとって判断基準の形成や一貫性の維持がより重要となる。

Bian et al. (2022) は、勝者や合格者数が限定される競争や、評価基準の一貫性が強く求められる状況において、初期の評価が後続の判断に大きな制約を与えるため、評価者が序盤で慎重に基準を設定しようとし、キャリブレーションが強く働くことを指摘している。本研究で観察される、参加者数の少ない大会や G 級以上の大会において順序効果が大きいという結果は、初期評価の相対的重要性や基準形成の困難さが高まる環境ほどキャリブレーションが強まるという点で、同論文の議論と整合的である。

どうして順序効果が生じるのか

次にどうして順序効果が生じるのかを検討する。今回のピアノコンクールの設定において、先行研究で既に指摘されている「疲労」や「動機」といったメカニズムに当てはまるものがあるだろうか。

結論から述べると、本研究の設定において、審査員の疲労が順序効果を生じさせているという証拠は見当たらない。もし審査員の疲労が原因ならば、今回のコンクールの予選のように、各大会・級における不合格者が（合格者よりも）多い場合は、コンクール開始から時間が経過するほど、疲労によって不合格というデフォルトの決定が増えると予測される。しかし、図 2 を見ても、時間が経過している後ろの演奏順の方が不利になるという傾向はみられない。また、表 4 における最終演奏者の係数は統計的に有意ではない。加えて、9 人以下の小規模な大会・級のほうが順序効果が大きいことも疲労仮説と整合的ではない。

長時間の審査が本当に疲労を蓄積するかどうかを確認するため、株式会社日立システムズと疲労科学研究所と共同開発した疲労測定器を用いて、大会当日の審査員の疲労を直接計測した（詳細は、Appendix 4 および Appendix 5 を参照）。この測定器は指先から心電波・脈波を計測できる自律神経測定器を用いて、心電波、脈波を計測することができる。2023 年度の本選において、106 名の審査員に、測定器を用いて、午前と午後に少なくとも 1 回の測定を依頼した（1 回につき、1 人約 2 分程度で測定が可

能で、のべ約 370 回の計測が行われた)^{4,5}。これをみると、コンクールの開示時刻から、時間の経過とともに審査員の疲労が顕著に蓄積しているというエビデンスは得られない。ただし、この測定機は身体的・精神的な疲労を計測することができたとしても、認知的な疲労を計測することが出来ていない点には注意が必要である。

次に「動機」である。既に述べたように、審査員が審査の一貫性を保ちたいと考える動機付けは、(1) 戦略的一貫性、(2) 割り当て、(3) キャリブレーションという 3 つがあり、それぞれ異なる結果をもたらす。まず、(1)の戦略的一貫性が生じる可能性は高くない。なぜなら、本研究で分析の対象とするコンクールの設定では、いったん審査が終了した後で、第三者が検証することや演奏者が演奏をやり直すことは許されていないためだ。もっとも、評価者が完全に外部からの検証を想定していなくとも、内的な一貫性を保ちたいという欲求が判断に影響する可能性は否定できないが、本研究の制度的設定においては、その影響は限定的と考えられる。

一方、(2)の割り当てが生じる可能性はある。本コンクールにおいては上位 35～45%を次選に出場させるという明示的な割り当てがある。しかし、割り当てを意識した意思決定が行われる場合、序盤よりも後半のほうが評価が厳しくなると予測され、表 4 の結果とは矛盾する。

更に、Chen et al. (2016)に倣って、割り当ての影響を統計的に検証したのが Appendix 6 である。割り当ての理論的根拠に基づくと、意思決定者にとって重要なのは、これまでに許容された決定の割合のはずであり、もし割り当てが順序効果が生じる主要な要因であれば、決定の割合を制御した後、順序 (Recency) の効果は消失するはずである。しかし、Appendix 6 で示されるように、実際には、決定の割合や審査員の属性を制御した後も、審査員が直前の決定に対して強く反応することが示されている。つまり、審査員の決定が、割り当てではなく、直前の決定の「順序」自体に影響されていることを示されており、割り当てが順序効果の主要なメカニズムである可能性は低いことを示唆している。直前演奏者の標準化得点との相関は正で統計的に有意となっていることから、評価の反転を予測するギャンブラーの誤謬によるものとも言えない。

⁴ 疲労やストレスの評価方法としては、尿検査、血液検査、唾液検査などがあり、血液検査の精度が高いと言われるが、血液検査は被験者の負担も重く、費用もかかるため、容易ではない。このため、自律神経のバランスや活動量を見る方法が用いられることが多い。同測定器は既に複数の研究で利用実績がある (Kume et al. 2017; Tamada et al., 2018; Mizuno et al., 2017)。

⁵ 疲労の計測は、本研究で分析の対象としている予選ではなく、予選の後に行われる本選で行われた。その理由は、予選と本選は審査員が重複しているため、審査員を対象にするのであれば予選でも本選でも変わらないということに加え、2023 年度は予選で実験を行ったため、予選で疲労計測を行った場合、実験の結果に影響を与える可能性を排除するためである。

(3)のキャリブレーションはどうだろうか⁶。Chen et al. (2016) はキャリブレーションが働くかどうかを確認するために、移民裁判所の審査データを用い、その日の評価対象が1件のみのケースと、複数件の評価の1件目に当たったケースを比較している。理論的予測としてはキャリブレーションが働くならば、後のケースと比較して一貫性を保つ必要がないので「その日の唯一のケース」では、極端な評価（Chen et al. (2016) では寛大な決定）を避けようとする動機が働かないため、「複数ケースの1件目」よりもむしろ極端な評価（寛大な決定）が多くなる可能性があるためだ。

これに倣い、本論文で用いたデータを用いて、演奏者1人のみの大会・級と複数人の大会・級で演奏順1番目の得点および予選通過率を比較した。なお、演奏者が1人のみの大会・級とそれ以外の大会を比較してみると、審査員の属性には差がないものの、出場者や大会の属性には統計的に有意な差があるため（Appendix 7）、まずは単純に演奏者1人のみの大会と複数人の大会で演奏順1番目を比較し、有意差のある変数をコントロール変数として追加していくことで、結果の頑健性を確認した。その結果を見てみると、演奏者1人のみの大会・級では、演奏順1番目の評価と通過率が大きく高くなっており、標準化スコアが約 $0.11\text{ }s.d.$ も、予選通過率も約 +34~35 ポイント高くなっている（Appendix 8）。これらはいずれも、年・級・地区固定効果や共変量を加えても頑健で、1%水準で統計的に有意である。Chen et al. (2016) で示されたのと同様に、演奏者が1人の大会では、極端な評価を避けようとする動機が働いておらず、評価が単独で行われるか否かが判断に影響を与えていることがわかる。

Arai & Okazawa (2025) は1日に2回開催されるお笑いコンテストのうち、1回目と2回目のコンテストの順序効果を比較し、2回目の方が順序効果が小さいことを発見している。1回目の審査で、出場者のパフォーマンスの分布を知ることになり、キャリブレーションが進むためである。このことを踏まえ、本研究では、予選の後に、予選通過者だけを集めて行われる本選の順序効果を予選と比較することにした。結果を見てみると、本選と予選の順序効果にはほとんど差がないことがわかる（Appendix 9）。9人以下と10人以上の大会・級、F級以下とG級以上の大会・級を比較しても、予選との間に大きなパターンの違いは見られない。1日に2回のコンテストが行われる Arai & Okazawa (2025) の設定と比較すると、本研究におけるピアノコンクールの本選と予選の間には数か月の期間があり、本選と予選の間で審査員の重複はあるも

⁶ Arai & Okazawa (2025) は日本のテレビ番組「爆笑オンエアバトル」（NHK）におけるコンテスト形式のデータを用い、最初に登場することが有利であることを明らかにしている。著者らはこれがキャリブレーションによって説明されると言う。Bian et al. (2022) のように、最初に登場することは不利になるという研究もある中、このケースでは10組中5組が勝者になるという合格率の高い競争であり、評価者のデフォルト行動は「勝者に投票する（肯定的な評価）」であることが最初に登場することが有利になる理由であると説明している。つまり、キャリブレーションが働く場合、最初の候補者に不利に働くか、有利に働くかは状況依存的であると言える。

の、全員が同じというわけではない。何よりも演奏者のパフォーマンスの質の分布が大きく異なっている。このため、Arai & Okazawa (2025) の設定と比較すると、本選の審査員が、予選の情報を元に出場者のパフォーマンスの分布を正確に把握できるとは限らないため、このような結果になっていると解釈できる。この結果は、キャリブレーションが短期間・同一文脈で繰り返される場合には進行しやすい一方、評価対象の分布や審査環境が大きく変化する場合には、必ずしも順序効果を緩和するとは限らないことを示唆している。

総じて、本研究の設定ではキャリブレーションが順序効果の主要なメカニズムとして働いている可能性が高い。一方で、予選と本選の順序効果に大きな差が見られないことから、評価経験の蓄積が必ずしも順序効果を緩和するとは限らないことも示唆される。加えて、観察データに基づく分析である以上、複数の動機が同時に作用している可能性を完全に排除することはできない点には留意が必要である。

順序効果を緩和できるか：審査員の経験年数とリハーサルの異質性

順序効果を緩和するためにどのような要因が有効であるのかを検討するため、本節では審査員の経験および審査開始前のリハーサルに着目し、2004～2022 年度の観察データを用いて分析を行う。先に述べたように、順序効果は最初の 3 番目までが顕著であることから、演奏順が 1～3 番目までの演奏者に対する順序効果について分析を進める（これは、後に述べるとおり、2023 年度のデータを用いて、演奏順が 1～3 番目の奏者についてのプライミングを行ったこととも関係している）。

表 5 は順序効果の審査員の経験の異質性をみたものである。審査員の経験をあらわす変数として、次の 5 つを用いた。審査員自身がピアノ指導者としてどの程度経験があるか、指導者賞の受賞回数、ピアノ指導者ライセンスを保有しているかどうか、セミナー講師かどうか、プロのピアニストかどうか、である。これらの変数と演奏順の交差項を見てみると、指導者受賞回数が多い場合、わずかに順序効果を軽減することがわかるが、係数は極めて小さい。ピアノ指導者ライセンス、セミナー講師、プロのピアニストの変数との交差項は統計的に有意ではない。Appendix 3 で審査員の経験年数をより細かく分けて行った推定をみても同様である。つまり、審査員の経験年数や資格、専門性が高い場合であっても、順序効果が緩和されるという証拠は得られなかった。これは、順序効果が単なる未熟さや判断能力の不足によって生じているわけではなく、経験豊富な評価者であっても回避が難しい構造的な問題である可能性が示唆される。

次に、リハーサルがどのような効果を持つのかということにも着目した。Bian et al. (2022) は、順序効果を軽減する方法として、審査の開始時に評価対象外のパフォーマ

ンスを含めることで、評価者に全体の質の分布に関する情報を与えるという方法を理論的に提案している。本研究の対象としたピティナ・ピアノコンペティションでは、A2～F 級で、希望者を対象として、大会開始前にリハーサルを行う。過去のデータを見ると、およそ半数の大会でリハーサルが行われているため、これを利用して Bian et al. (2022)の仮説を実際に検証した。

ここで注意が必要なのは、このリハーサルは、コンサートなどで一般に行われる「練習」ではないという点である。ピティナ・ピアノコンペティションにおけるリハーサルは、通常の参加と同じ参加資格、料金、会場、条件で実施される。審査員は本番と同様に評点を付与するが、本選への選出や賞の対象とはならない。このため、リハーサルは審査員に対して、当該大会・級におけるパフォーマンスの質の分布に関する情報を部分的に提供する可能性がある。

表 6 では、2 列目でリハーサルも演奏順に加えた結果を、3 列目ではリハーサルがない大会に絞って演奏順の効果をみている。4 列目ではリハーサルのある大会に絞って、リハーサルを含む演奏順の効果を、5 列目ではリハーサルのある大会に絞って、リハーサルを除く演奏順の効果を見ている。いずれも、1 列目のベースラインの結果とほとんど差がなく、リハーサルがあることが順序効果を緩和するという明確なエビデンスは得られない。ベースラインと比較した時に係数に統計的に有意な差があるかどうかを t 値は各演奏順の係数の差の 2 標本 Wald 検定で確認したところ、リハーサルのある大会に絞って、リハーサルを除く演奏順の係数 (-0.087) との差だけが 5% 水準で統計的に有意な差があった。つまり、リハーサルの実施は、審査員に一定の情報を与える可能性があるにもかかわらず、順序効果を緩和する効果を持たないことが示された。むしろ一部の推定では、リハーサルが行われた大会において順序効果が大きくなる傾向も観察されており、部分的な情報ではキャリブレーションを促進するとは限らないことが示唆される。こうした観察データに基づく分析結果を踏まえ、次節では、2023 年度のコンクールにおいて実施したランダム化比較試験により、順序効果を直接的に是正する介入の因果効果を検証する。

順序効果を緩和する介入：フィールド実験

本研究では、特に評価者が逐次評価の初期段階で基準を過度に保守的に設定してしまう「キャリブレーション」に基づく順序効果に着目し、それを是正できるかを検証する。キャリブレーションによる順序効果は、評価者が逐次評価の初期段階において、パフォーマンスの質の分布に関する情報を十分に持たないために、極端な評価を避け、基準を過度に保守的に設定してしまうことから生じると考えられている

(Unkelbach & Memmert, 2014; Bian et al., 2022; Plonsky et al., 2023)。そこで、過去

の同一コンクールのデータに基づいて、序盤の評価が体系的に低くなる傾向と、それが合否に影響しうることを具体的な数値で示すことで、評価開始前に分布に関する事前情報を与える。Page & Page (2010)は順序効果の存在を評価者に認識させることは順序効果を緩和する可能性があるが、こうした検討は未だなされていないと指摘している。

今回の実験では、これまでの手続きや運用を大きく変えることなく実施できるように、簡易な紙のチラシと自身のスマートフォンで見られる動画によって審査員への情報提供を行い、その効果をクラスターランダム化比較実験によって評価した⁷。チラシでは、過去のデータを用いて、演奏順が1～3番目の奏者とそれ以降の奏者との評点差を級ごとに示すとともに、その差が予選・本選の合否にどの程度影響しうるかを、合格最低点と不合格者最高点との差を用いて具体的に示した（全文は Appendix 7）。

このことを5分程度でわかりやすく説明した動画のQRコードも添付した。この中では、ピアノ以外にもお笑いコンテストやスポーツの演技の評価においても順序効果が生じていることなども紹介した。この動画の内容とスクリプトは Appendix 11 に掲載した。当然、これらの情報提供は演奏者には知らされず、演奏者や演奏する環境は一切影響を受けない。

これは、情報提供実験（Information Provision Experiments, IPEs）と位置付けられる。経済主体の意思決定は選好・制約・信念で特徴づけられるが、実証的に信念を操作する手段として急速に普及しつつある。しかし、過去の研究によれば、様々な介入は、信念を変更させる効果は大きい、行動への効果は小さいことが多い（Haaland, et al., 2023）。0.15 *s.d.*の効果を検出するために80%の検出力では、処置群に割り当てられる被験者は700人程度が必要となる。

また、参加者の事前信念からどれだけ乖離した新情報を与えるか、乖離が大きいほど被験者の「ショック」が強くなることがわかっている。しかし、参加者がその情報を「信頼できる」と感じる程度も重要であり、これは情報源や提示方法で左右され

⁷ 介入のプランを考える際には、コンクールの現場でどのようなリソースが利用可能か、審査の手続きや運用を変更できるか、と言う視点は重要である。過年度のデータをもとに評価の順序を最適化するという提案もあるが（Akl & Tewfik, 2016）、今回の設定では35%程度の演奏者が初出場であり、過年度のパフォーマンスについてのデータが利用可能ではない。アルゴリズムに基づいて事後的に逐次バイアスを修正した後のスコアを算出するためには、審査員にICT機器などを用いて評価を入力してもらう必要があるが、今回の設定では年配の審査員も多く、ICT機器の使用に慣れていないため、ICT機器の導入そのものが実験の結果に影響を与えてしまう交絡要因となってしまう恐れもある。これら以外にも、Unkelbach & Memmert (2014)やBian et al. (2022)は、順序効果を軽減させるために、一連のパフォーマンスを見終わってからまとめて評価する「後置評価方式」や、一気に複数人を連続評価するのではなく、演奏ごとに休憩や時間間隔を設けるなどして、連続的な判断を意識させない方法を提案しているが、こうした方法を導入することはコンクールの運営方式や評価方法をかなり大きく変化させてしまい、介入の効果と運営方式や評価方法を変更したことの効果の識別を難しくしてしまう。

る。この「ショック」と「信頼度」にはトレードオフがあり、例えば、既に知っている情報と大差がなければ、ショックが不足することで情報介入の効果が限定的となってしまう。

このため、過去のピティナ・ピアノコンペティションのデータを用いることで信頼性を高める一方、演奏順によっては本来合格していたはずの奏者が、合格できなかったという不運を生んでいることを強調したり、ピアノ以外でも「評価のバイアス」が生じており、社会における幅広い文脈において重要であるということを伝えることで、「ショック」が大きくなるように工夫した。

チラシの配布は、大会が開始になる時間（多くの場合、午前9時頃）の直前に、審査員同士の顔合わせと簡単な打ち合わせが行われる際に、審査員長によって行われた。およそ10～15分程度をかけて、チラシの内容の説明が行われた。

割り付け

本研究では、審査員間での情報共有によるコンタミネーションを防ぐため、大会の開催日を単位としたクラスターランダム化を行い、チラシと動画による情報提供の有無を割り付けた。まず、実験を行った2023年度の予選について述べる。2023年度にはコロナウィルス感染症以前と同程度の254大会が開催された。5月1日時点で開催日が未定の2大会、5月下旬に開催された10大会、8月に開催された3大会を除く231大会（級・大会別では1,470）を分析の対象にした⁸。この231大会には、のべ23,427名の演奏者が参加し、のべ1,161名の審査員が評価を行った。1大会あたりの平均参加者数は19.3人であり、高校3年生以下が出場するF級以下が全体の97.7%を占める。

割付には開催日の情報を用いた。ほとんどの大会は6～7月の間に実施され、7回の金土日及び祝祭日に集中している。この7回の開催日を事前に乱数を用いて介入群と対照群に割り付けた。開催日別の大会数は、図4に示した。介入群に割り当てられた大会は47.6%である。

大会の開催日付が疑似ランダムであることを示すため、大会会場および日程が決定されるプロセスについて説明する。まず大会の運営者は開催の前年に大会会場の予約を始める。3月1日に日付と会場が公表され、4月1日に参加の申し込みが始まる。ピアノコンクールを実施するためには、防音設備を備え、楽器演奏が可能な会場である必要がある。会場利用が抽選で決定されることもままたり、会場の確保は実務的に

⁸ 他の大会は土日及び祝祭日に行われているのに対し、5月に実施された予選の10大会は平日に開催されていることから、実験の対象から外すこととした。また8月に開催された3大会はソロではなくデュオの大会のため、実験の対象から外した。

容易ではない。過去のデータを見ても、同一大会が前年と同日または同週に開催される例は少なく、事前に開催会場における正確な開催日を予測することは困難である。このような制度設計の下では、応募者が特定の開催日を狙って、自宅から遠い大会を選択する余地は限定的であり、開催日を用いた割り付けは、少なくとも観測可能な要因に関しては疑似ランダムであると考えられる。また、開催日単位での割付を採用することで、同一日に審査を担当する審査員間での情報共有や行動の相互影響によるコンタミネーションを最小限に抑得るという目的もある。

審査員が複数の大会で審査を担当することもあるため、同一審査員が複数回処置を受ける可能性がある。このため、介入頻度が極端に偏らないよう、割付確率を事前に検討した。過去（2004～2022 年度、2020 年度を除く）の審査員の担当回数分布を用いてベルヌーイ試行を行い、処置確率を 0.5 とした場合に、審査員が処置を受ける回数の分布がどのようになるかを事前に確認した。処置を受ける確率 p が 0.5 の場合、処置回数が 1 回の審査員は 46.48%、2 回になる審査員の割合が 18.05%、3 回以上になる審査員の割合が 4.80% となった。

表 8 は処置群と対照群を比較したバランステストの結果である。2 群の差の検定を行うと、大会、審査員の属性については両群の間に差はみられないが、演奏者については一部の変数で統計的に有意な差が見られる。これらは割付後に決定されるものではないため、偶然によるものと考えられるが、共変量として統制することで推定結果に与える影響を調整している。

フィールド実験の分析結果

本節では、順序効果が評価者のキャリブレーションに基づいて生じている場合に、順序効果の存在を明示的に認識させる情報提供がその軽減に有効かどうかを、2023 年度のクラスターランダム化比較実験を用いて検証する。表 9 では、アウトカムを演奏者ごとの標準化得点として、介入の因果効果を推定した。欠席者・リハーサルを除く演奏者の大会・級別の演奏順のうち、1～3 番目を 1、それ以外を 0 としたダミー変数と介入を受けたかどうかのダミー変数の交差項の係数が介入の因果効果を示す。1 列目は、ベースラインの推定である。2 列目は、これに加え、エントリー回数、リハーサルエントリー回数、表 8 で 2 群の間に統計的に有意な差がある変数に加え、大会規模（欠席・リハ除く最終演奏者の演奏順）をコントロールし、級、審査員、大会開催地区、累計エントリー回数の固定効果も加えた。3 列目はそれに加え、リハーサルのある大会であれば 1、それ以外を 0 としたダミー変数のみをコントロールした。すべて大会ごとに標準偏差をクラスターリングしている。これ以外にも頑健性を確認するため、リハーサルのある大会とない大会を分けて推定した結果も掲載した。加えて、

頑健性を確認するため、予選を通過して本選に進んだかを表すダミー変数、予選不通過者のうち上位に与えられる予選奨励賞を受賞したかをあらわすダミー変数をアウトカムとした推定も行った（表9の参考）。

結果を見てみると、いずれの推定においても、演奏順ダミーと処置群ダミーの交差項の係数は負で統計的に有意ではない。この結果は、情報提供によって順序効果の存在を認識させること自体はできても、評価行動の修正には必ずしも直結しない可能性を示唆している。

次に、表10で大会・級の属性による異質性を確認する。この推定は、表9の2列目と同じ定式化で推定している。1列目は、介入の強度の影響を見るため、2回以上の介入を受けた審査員の異質性を見ている。2列目は、記憶が新しいことの影響を見るため、情報提供の直後に演奏が始まった午前の1グループ目の異質性も推定した。これらの三重交差項の係数はいずれも統計的に有意ではなかった。

更に、表10の3列目では、表4で順序効果の係数が大きかった9人以下の小規模な大会・級における異質性を見ている。これも演奏順と介入群であれば1を取るダミー変数、9人以下の大会・級であれば1を取るダミー変数の三重交差項の係数は有意ではなく、9人以下の小規模な大会・級で順序効果を軽減したという証拠は得られない。次に表4で順序効果の係数が大きかったG級以上の上級者が参加する大会・級における異質性を見ている。三重交差項の係数は0.271で5%水準で統計的に有意である。なお、G級以上における演奏順ダミーの係数は-0.471（標準誤差0.078）であり、介入によって順序効果のおよそ58%が軽減されたことになる。G級以上の大会は、評価対象となるパフォーマンスの質が高く、かつ上位進出の可否が将来のキャリアに大きな影響を与える高ステークスな競争環境である。このような状況では、評価者が序盤で慎重に評価基準を設定しようとする動機が特に強く働くと考えられる

（Bian et al., 2022）。この結果は、高ステークスな環境においては、評価者が自身の判断の歪みを意識することで、キャリブレーションに基づく過度に保守的な初期評価を部分的に修正する可能性があることを示唆している。一方、F級以下ではもともとの順序効果が比較的小さく、評価者の判断基準も比較的安定していると考えられるため、情報提供による追加的な行動修正の余地が限定的であった可能性がある。

念のため、G級以上とF級以下にサンプルを分けた推定も掲載した。サンプルサイズが小さくなるため、G級以上の演奏順ダミーと介入ダミーの交差項は統計的に有意ではないものの、係数は0.267（標準誤差0.156）で、やはり約50%は軽減することができている。このため、もともと順序効果が大きいG級以上では、一定の介入効果が見られることが明らかになった。Appendix 10では審査員の経験による異質性を確認しているが、異質性は確認できなかった。また、予選通過率および予選奨励賞をア

ウトカムとした分析も行ったが、異質性は確認されなかった⁹。

実験終了後に、審査員に対して質問紙調査を行い、予選に審査員として大会に参加した 329 名から回答があった（2023 年度の予選の審査員数 352 名、回収率 93.4%）。

大会期間中に「演奏順がパフォーマンスにもたらす影響」についての 1 枚のチラシを何回見たかという問いについては、1 回が 28.9%、2 回が 30.7%、3 回以上が 23.4%、見ていない（一度も介入を受けない純粋な対照群）が 12.8%となっている。一方、動画については 21.3%が視聴したと回答しており、実際に大会終了後に確認した動画視聴回数のユニークな IP アドレスの 78 回（22.1%）とほぼ一致している。

チラシや動画以外に、審査員同士の会話の中で、情報提供の内容が知られることでスピルオーバーが生じる可能性があるため、大会期間中に、「演奏順がパフォーマンスにもたらす影響」について審査員同士で話をする機会があったかを尋ねたところ、18.2%が「全くなかった」と回答しており、このように回答した審査員は大会開催中にチラシを「見ていない」と回答している審査員（一度も介入を受けない純粋な対照群）と 59.5%が一致している。なお、今回の大会期間中に「最終奏者まですべて採点し終わった後で、演奏順が最初の奏者の採点を付けなおしたことがあったか」という問いについては、24.3%が「付けなおした」と回答しており、この全員が 1 回以上介入を受けた審査員である。この結果を見る限り、介入が審査員の行動に影響を与えなかったとは言えない。

自由回答を見てみると、「確かに演奏順が最初の奏者については控えめに点数を付けていた」とか「後に良い演奏が来るかもしれない、と始めの方の採点を控えめに付ける審査員は確かに少なくない」という主旨のコメントは複数見られており、順序効果があることを自覚していた審査員は多いことがわかる。そして、このように最初の奏者が不利になる状況が生まれるのは、「ピアノの弦がまだ温まっていない」とか「（審査員の）耳がホールの響きに慣れていないから評価が定まらない」というピアノの調律や会場の音響などの技術的な問題だと捉えている審査員が多かったこともわかった。そして「今年の結果は、例年より 1 番や前半に高得点がつけられている事が多かった気がする。もしかして、この調査で意識的に審査員の心理状態が変わったのかなとも推測した。」というコメントもあり、中には介入によって結果が変わったことを期待する声もあった。しかし、こうしたアネクドタルな審査員の心証とは異なり、実証分析の結果からは、全体として順序効果が緩和されたという確かな証拠は得られない。

本研究の介入は、こうした「序盤で極端な評価を避ける」という内生的な基準調整

⁹ この分析は紙幅の制約により掲載しないが、要請があれば提供可能である。

が、結果として初期の演奏者を不利にしていることを明示的に認識させるものである。質問紙調査の結果からは、介入によって審査員の認識や内省が促されたことは確認できるが、その効果が行動として一貫して現れ、順序効果が修正されるのは、高ステークスで、評価者が序盤で慎重に評価基準を設定しようとする動機が特に強く働く状況に限られることも明らかになった。以上の結果から、情報提供は全体として順序効果を解消するものではないが、順序効果が特に強い状況では部分的に有効であることが示された。逐次評価の公平性を確保するためには、注意喚起や情報提供、意識改革ではなく、評価手続きそのものの設計を見直す構造的アプローチが必要な可能性がある。

Conclusion

本研究は、日本最大規模のピアノコンクールにおける長期の観察データとクラスターランダム化比較試験（RCT）を組み合わせることで、逐次評価における順序効果について検証した。とりわけ、本研究は、同一の制度設定の下で観察データとフィールド実験を組み合わせることで、順序効果の存在を確認することから、その緩和を企図した介入の効果検証までを一貫して行った点に特徴がある。逐次評価は、音楽・スポーツのコンテスト、採用面接、助成金審査など幅広い場面で用いられているが、現実社会に即した大規模データとフィールド実験の双方を用いて検討した研究は極めて少ない。本研究は、そのギャップを埋めることを目的とした。

本研究で明らかになったことは下記の通りである。第一に、2004～2022年度の観察データを用いた分析から、本コンクールにおいては一貫して「前の演奏順が不利になる」順序効果が存在することが明らかになった。例えば、1番最初の演奏者は、11番目以降の演奏者と比較すると $0.117\text{ }s.d.$ も評点が低く、こうした不利は3番目までの演奏者に特に顕著である。さらに、この順序効果は大会・級の属性によって大きく異なる。参加者数が少ない大会や、G級以上といった演奏者のパフォーマンスの質が高い大会において、順序効果はより大きくなることが示された。一方で、審査員の経験年数や資格、リハーサルの実施といった要因が順序効果を一貫して緩和するという証拠は得られなかった。

第二に、順序効果が生じるメカニズムについて検討した結果、本研究の設定では、評価者が序盤で評価基準を過度に保守的に設定してしまう「キャリブレーション」が、順序効果を説明する有力な要因の一つである可能性が示唆された。審査員は演奏者全体のパフォーマンス分布を十分に把握できない序盤において、評価の整合性を保とうとする動機から、極端な評価を避け、結果として最初の演奏者に厳しい評価を与える傾向がある。演奏者が1人しかいない大会ではこの不利が観察されないことや、

競争が厳しい G 級以上で順序効果が大きいことは、この解釈と整合的である。

第三に、2023 年度のデータを用いて実施したクラスターランダム化比較試験では、順序効果を緩和することを目的とした情報提供介入（プライミング）の因果効果を検証した。審査員に対して、過去データに基づく順序効果の存在やその大きさを具体的に提示することで、最初の 1～3 番目の演奏者に対する不利が軽減されるかを検討したが、全体としては順序効果が有意に緩和されたという強い証拠は得られなかった。

しかしながら、異質性分析からは重要な示唆も得られた。演奏者のパフォーマンスの質が高い G 級以上の大会では、介入によって順序効果がおおよそ半分程度に緩和される可能性が示された。この結果は、評価基準の設定が難しく、かつ初期評価の影響が大きい高ステークス環境においては、評価者が自身の判断の歪みに自覚的になることで、キャリブレーションに基づく順序効果を部分的に修正し得ることを示唆している。一方で、その調整は順序効果全体の 50～60% 程度を相殺するにとどまっており、順序効果を完全に解消するには至っていない。

最後に、本研究にはいくつかの限界も存在する。第一に、観察データからは評価者の内的な認知過程を直接観測することはできず、キャリブレーション以外の動機が完全に排除されたわけではない。第二に、介入の効果は特定の設定や情報提示方法に依存しており、異なる設計やより強度の高い介入によって結果が異なる可能性も残されている。

逐次評価は、教育、労働、市場取引、司法など多くの分野で不可欠な制度である。本研究の結果は、逐次評価が広く用いられている現実社会—採用面接、助成金審査、学術評価、スポーツや芸術のコンテストなど—に対して重要な含意を持つ。一方で、2023 年度のフィールド実験の結果は、順序効果が評価者の認識不足のみに起因するものではなく、逐次的に判断を下さざるを得ない制度そのものに内在する問題である可能性が高いことを示している。順序効果を完全に解消するためには、評価者個人の注意や意識改革に依存するのではなく、評価順に応じたアルゴリズムによる事後的な補正を行うなど、制度的な対応を検討する必要がある。

Reference

- Akl, E. A., Maroun, N., Nehmé, R., Schünemann, H. J., & Tewfik, T. L. (2014). Optimal information sequencing for cognitive bias mitigation. *Judgment and Decision Making*, 9(6), 485–493.
- Akl, E. A., El-Khoury, C., Nehmé, R., & Tewfik, T. L. (2016). Designing interventions to mitigate cognitive biases in human decisions. *Frontiers in Psychology*, 7, 1402.
- Antipov, E. A., & Pokryshevskaya, E. B. (2017). Order effects in online consumer reviews: An experimental study. *Electronic Commerce Research and Applications*, 22, 1–11.
- Arai, R., & Okazawa, R. (2025). *Is it advantageous to be first? Evidence from a TV comedy program*. *Journal of Economic Behavior & Organization*, 234, 107009.
- Asmat, R., Borowiecki, K. J., & Law, M. T. (2023). Do experts and laypersons differ? Some evidence from international classical music competitions. *Journal of Economic Behavior & Organization*, 214, 270–290.
- Bhargava, S., & Fisman, R. (2014). Contrast effects in sequential decisions: Evidence from speed dating. *Review of Economic Studies*, 81(2), 531–558.
- Bian, X., Wang, X., & Xu, H. (2022). *Information calibration and evaluation bias in sequential contests*. *Management Science*, 68(9), 6664–6682.
- Bruine de Bruin, W. (2005a). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, 118(3), 245–260.
- Bruine de Bruin, W. (2005b). Who judges best? Individual differences in accuracy and bias in comparative judgment. *Organizational Behavior and Human Decision Processes*, 96(1), 68–81.
- Bruine de Bruin, W. (2006a). Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, 123(3), 299–311.
- Bruine de Bruin, W. (2006b). Judge mental effort in sequential choice. *Organizational Behavior and Human Decision Processes*, 99(2), 113–129.
- Carney, D. R., & Banaji, M. R. (2012). First is best. *PloS one*, 7(6), e35088.
- Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *Quarterly Journal of Economics*, 131(3), 1181–1242.
- Clingingsmith, D., & Shane, S. (2017). Let others go first: Order effects in sequential contests. *PloS one*, 12(8), e0183615.
- Collins, A., McKenzie, J., & Williams, L. V. (2019). When is a talent contest not a talent contest? Sequential performance bias in expert evaluation. *Economics Letters*, 177, 94–98.
- Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17), 6889–6892.
- Falk, A., & Zimmermann, F. (2018). Information processing and commitment. *Economic Journal*, 128(613), 1983–2002.
- Finkelstein, S., & Hambrick, D. C. (1990). Top-management-team tenure and organizational outcomes: The moderating role of managerial discretion. *Administrative Science Quarterly*, 484–503.
- Flôres, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(1), 97–104.
- Ginsburgh, V. A., & van Ours, J. C. (2003). Expert opinion and compensation: Evidence from a musical competition. *American Economic Review*, 93(1), 289–296.
- Glejser, H., & Heyndels, B. (2001). Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth Music Contest. *Journal of Cultural Economics*, 25(2), 109–129.
- Haaland, I., Roth, C., & Wohlfart, J. (2023). Designing information provision experiments.

- Journal of Economic Perspectives*, 37(2), 201–226.
- Haerem, T., & Rau, D. (2007). The influence of degree of expertise and objective task complexity on perceived task complexity and performance. *Journal of Applied Psychology*, 92(5), 1320.
- Hartzmark, S. M., & Shue, K. (2018). A tough act to follow: Contrast effects in financial markets. *Journal of Finance*, 73(4), 1567-1613.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1), 1-55.
- Ibanez, M. R., & Toffel, M. W. (2020). How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science*, 66(6), 2396-2416.
- Kume, S., Nishimura, Y., Mizuno, K., Sakimoto, N., Hori, H., Tamura, Y., ... & Kataoka, Y. (2017). Music improves subjective feelings leading to cardiac autonomic nervous modulation: a pilot study. *Frontiers in Neuroscience*, 11, 108.
- Leibovitch, A. (2016). Relative judgments. *Journal of Legal Studies*, 45(2), 281-330.
- Mantonakis, A., Rodero, P., Lesschaeve, I., & Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science*, 20(11), 1309-1312.
- Mizuno, K., Sasaki, A. T., Ebisu, K., Tajima, K., Kajimoto, O., Nojima, J., ... & Watanabe, Y. (2017). Hydrogen-rich water for improvements of mood, anxiety, and autonomic nerve function in daily life. *Medical Gas Research*, 7(4), 247-255.
- O'Sullivan, E. D., & Schofield, S. J. (2019). Cognitive bias in clinical practice. *Clinical Medicine*, 19(2), 131–134.
- Orazbayev, S. (2017). Sequential order as an extraneous factor in editorial decision. *Scientometrics*, 113(3), 1573-1592.
- Page, L., & Page, K. (2010). Last shall be first: A field study of biases in sequential performance evaluation. *Journal of Economic Behavior & Organization*, 73(2), 186–198.
- Plonsky, O., Chen, D. L., Netzer, L., Steiner, T., & Feldman, Y. (2023). Motivational drivers for serial position effects: Evidence from high-stakes legal decisions. *Journal of Applied Psychology*, 108(7), 1137.
- Quigley-McBride, A., Franco, G., McLaren, D. B., Mantonakis, A., & Garry, M. (2018). In the real world, people prefer their last whisky when tasting options in a long sequence. *Plos one*, 13(8), e0202732.
- Read, D., & Loewenstein, G. (1995). Diversification bias: Explaining the discrepancy in variety seeking between combined and separated choices. *Journal of Experimental Psychology: Applied*, 1(1), 34.
- Rothhoff, K. W. (2015). Sequential bias in judging gymnastics. *Journal of Sports Economics*, 16(6), 639–659.
- Simonson, I. (1990). The effect of purchase quantity and timing on variety-seeking behavior. *Journal of Marketing Research*, 27(2), 150-162.
- Simonsohn, U., & Gino, F. (2013). Daily horizons: Evidence of narrow bracketing in judgment from 10 years of MBA admissions interviews. *Psychological Science*, 24(2), 219-224.
- Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, 53(1), 1-26.
- Sunstein, C. R., Kahneman, D., Schkade, D., & Ritov, I. (2001). Predictably incoherent judgments. *Stanford Law Review*, 54, 1153.
- Unkelbach, C., & Memmert, D. (2014). Calibration of judgment in sequential evaluation: The serial position effect in sports judgments. *Journal of Experimental Psychology: Applied*, 20(1), 1–13.
- Unkelbach, C., Ostheimer, V., Fasold, F., & Memmert, D. (2012). A calibration explanation of serial position effects in evaluative judgments. *Organizational Behavior and Human Decision Processes*, 119(1), 103-113.

- Tamada, S., Ebisu, K., Yasuda, S., Kato, M., Ninomiya, N., Yamasaki, T., ... & Watanabe, Y. (2018). Kamikihito improves cancer-related fatigue by restoring balance between the sympathetic and parasympathetic nervous systems. *Prostate International*, 6(2), 55-60.
- Vives, M. L., Fernandez-Navia, T., Teixidó, J. J., & Serra-Burriel, M. (2021). Lenience breeds strictness: The generosity-erosion effect in hiring decisions. *Science Advances*, 7(17), eabe2045.
- Wang, J., & Pananjady, A. (2022). Modeling and correcting bias in sequential evaluation. arXiv preprint arXiv:2205.01607.
- 松山博幸. (2021). 審査のゆがみ: 全日本吹奏楽コンクールを例に. 応用経済学研究, 14, 45-68.

Figures and Tables

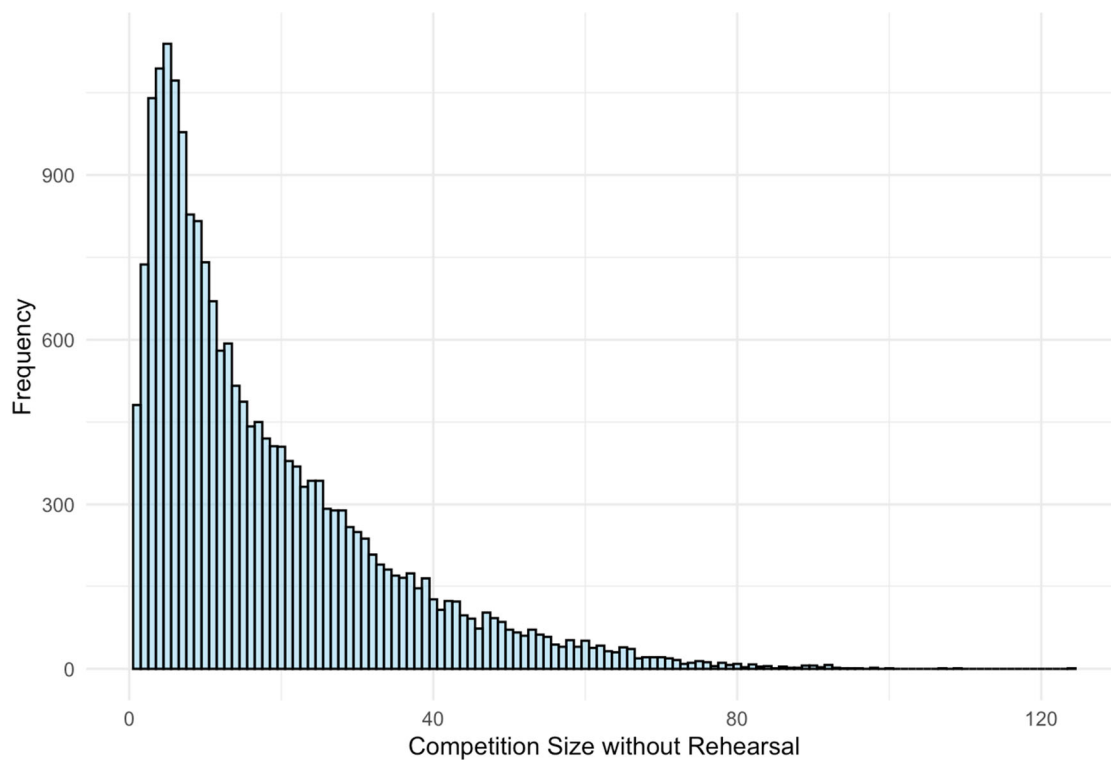


図1 大会・級の参加者数の分布（リハーサルを除く）

（注）2004～2022 年度（コロナウイルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティションに参加した演奏者のデータを用いて、大会・級の参加者数を示した。

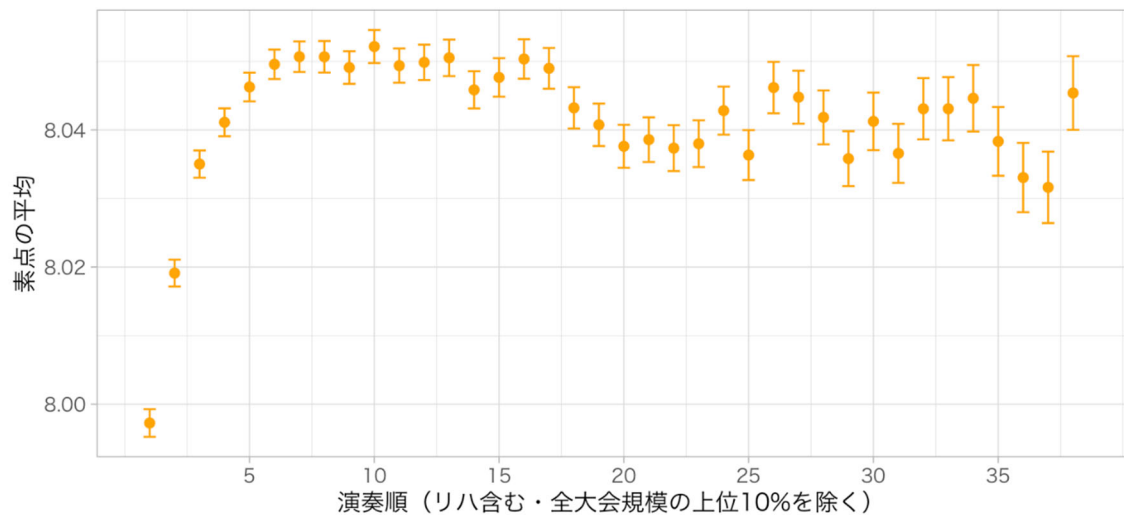
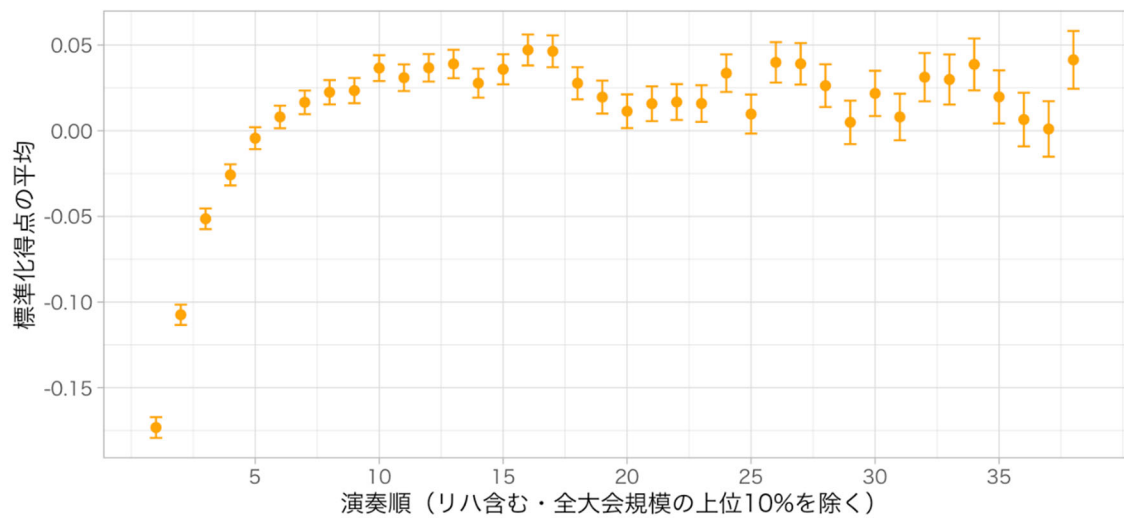


図2：演奏順ごとの評点

- (注) 1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティションに参加したのべ約 191 万人の演奏者の評点を用いた。標準化得点は、5 人の審査員の最高点と最低点を除いた平均点を、コンクールの開催年・地区・級ごとに平均 0、分散 1 となるように変換した値。
2. 演奏順は欠席者・リハーサルを除く。
3. 大会・級によって参加者数が異なるため、最終奏者までを含めると、最終奏者に近づくほど分散が大きくなるため、図では、各大会・級の最初の 90% までの演奏者の評点のみを図に含めた。

表 1 記述統計（2004～2022 年度）

大会の属性	Mean	SD	Max	Min	N
演奏者数（大会）	27.7	17.9	124	1	3,295
演奏者数（大会、リハーサル除く）	26.5	17.5	124	1	3,293
演奏者の属性					
年齢	10.400	3.190	79	2	394,096
性別（女性であれば 1）	0.829	0.376	1	0	393,781
大会エントリー回数	1.230	0.503	2	0	394,096
リハーサルエントリー回数	0.0733	0.289	7	0	394,096
初出場（初出場であれば 1）	0.287	0.452	1	0	394,292
審査員の属性					
指導者としての経験年数	17.700	8.710	45	0	3,969
性別（女性であれば 1）	0.750	0.433	1	0	4,105
ピアノ指導者ライセンス（ライセンスがあれば 1）	0.895	0.267	1	0	4,105
ピアノセミナー講師（講師であれば 1）	0.205	0.404	1	0	4,105
プロピアニスト（プロピアニストであれば 1）	0.073	0.260	1	0	4,105
指導者賞受賞回数	4.120	6.450	37	0	4,170

（注）1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティションに参加した演奏者のデータを用いた。

2. エントリー回数が 0 の奏者がいるが、これは本大会へのエントリー回数のみをカウントしているためで、当該年度・級でリハーサルのみに参加した演奏者についてはエントリー回数が 0 となる。また初出場は、（リハーサルも含めて）ピティナ・ピアノコンペティションの予選に初めて演奏した年・大会であれば 1、それ以外は 0 となるダミー変数。
3. 指導者賞は（1）予選を通過し本選に進んだ生徒が 8 名以上、（2）本選に 6 名以上の生徒が参加し、かつ、そのうち 2 名以上が本選優秀賞以上を受賞のいずれかの指導者に与えられる賞であり、指導者賞受賞回数はこれを受賞した回数、ピアノ指導者ライセンスは、ピアノ指導者またはピアノ指導者を目指しているものが受検する検定に合格した者、ピアノセミナー講師はピティナが主催、共催、後援するセミナーで講師を務めた経験がある者、を指す。

表 2 記述統計（2023 年度）

大会の属性	Mean	SD	Max	Min	N
演奏者数（大会）	19.30	14.20	65	1	231
演奏者数（大会、リハーサル除く）	19.20	14.30	63	1	225
演奏者の属性					
年齢	10.20	3.310	56	4	23,423
性別（女性であれば 1）	0.760	0.427	1	0	23,424
大会エントリー回数	1.330	0.484	1	0	23,427
リハーサルエントリー回数	0.032	0.201	4	0	23,427
初出場（初出場であれば 1）	0.347	0.476	1	0	23,427
審査員の属性					
指導者としての経験年数	22.40	10.60	47	2	1,157
性別（女性であれば 1）	0.777	0.416	1	0	1,161
ピアノ指導者ライセンス（ライセンスがあれば 1）	0.800	0.400	1	0	1,161
ピアノセミナー講師（講師であれば 1）	0.137	0.344	1	0	1,161
プロピアニスト（プロピアニストであれば 1）	0.058	0.233	1	0	1,161
指導者賞受賞回数	6.060	8.410	35	0	1,161

（注）2023 年度にビティナ・ピアノコンペティションに参加した演奏者と審査員のデータを用いた。各変数の定義は表 1 と同じ。

表3 演奏順のランダムネス

		大会・級別の参加者数で見た時 頻度が多い大会（2004～2022 年度）				大会・級別の参加者数で見た時 頻度が多い大会（2023 年度）			
		全体	5 人 (1 位)	4 人 (2 位)	6 人 (3 位)	全体	5 人 (1 位)	6 人 (2 位)	8 人 (3 位)
演奏者	年齢	0.000 (0.000)	-0.000 (0.001)	-0.001 (0.002)	-0.000 (0.001)	0.035 (0.069)	-0.509 (0.626)	-0.562 (0.464)	-0.297 (0.431)
	性別	0.001 (0.001)	-0.001 (0.013)	-0.018 (0.016)	0.007 (0.012)	-0.158 (0.482)	-1.842 (4.361)	1.105 (3.887)	2.384 (3.274)
	大会エントリー回数	-0.000 (0.001)	0.012 (0.012)	0.003 (0.014)	0.005 (0.011)	0.424 (0.519)	0.207 (4.043)	-0.833 (3.977)	1.458 (3.278)
	リハーサルエントリー数	0.001 (0.003)	-0.028 (0.035)	0.045 (0.037)	-0.035 (0.032)	-2.463 (1.746)	-6.844 (12.181)	2.812 (10.698)	-22.942** (10.594)
	初出場	-0.001 (0.001)	-0.008 (0.013)	-0.011 (0.016)	-0.006 (0.012)	-0.036 (0.546)	-9.390 (6.212)	-9.724** (4.780)	-2.803 (4.627)
審査員	経験年数	0.000 (0.000)	-0.000 (0.001)	-0.000 (0.002)	-0.000 (0.001)	0.001 (0.058)	-0.040 (0.481)	-0.099 (0.495)	0.023 (0.376)
	性別	0.001 (0.004)	-0.003 (0.041)	-0.001 (0.048)	0.001 (0.035)	0.193 (1.219)	-0.450 (10.812)	-0.132 (9.974)	-0.058 (9.429)
	ライセンス	0.000 (0.005)	0.006 (0.041)	0.005 (0.049)	0.001 (0.039)	0.022 (1.471)	1.531 (13.118)	2.257 (12.580)	-2.016 (9.830)
	セミナー講師	0.001 (0.003)	-0.002 (0.033)	-0.004 (0.038)	0.002 (0.030)	-0.116 (1.496)	1.177 (14.990)	3.036 (11.220)	-0.310 (10.523)
	プロピアニスト	-0.000 (0.006)	0.002 (0.056)	0.005 (0.068)	0.001 (0.051)	-0.043 (2.235)	1.493 (21.653)	-5.464 (16.477)	-1.490 (15.723)
	指導者賞受賞回数	0.000 (0.000)	0.000 (0.002)	-0.000 (0.002)	-0.000 (0.002)	0.005 (0.062)	0.007 (0.595)	-0.274 (0.627)	-0.036 (0.450)
N		370,460	5,680	4,364	6,422	22,632	415	474	616
Adj. R2		-0.000	-0.001	-0.002	-0.001	-0.000	-0.020	-0.013	-0.008
F 値		0.135	0.247	0.338	0.202	0.312	0.269	0.460	0.537

(注) 1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティションに参加した演奏者と審査員のデータを用いた。ただし、審査員の情報は、当該演奏者の審査を担当した審査員の平均値である。演奏順はリハーサルを除くパーセンタイル順位。各変数の定義は表 1 と同じ。F 値は固定効果を除く説明変数の同時有意性を年・大会開催地区クラスター頑健 Wald 検定で評価した。

2. ロバストネスチェックとして、級・大会別の参加者総数と大会の全参加者総数の上位の大会を取り出してバランステストを行った。

表 4 順序効果（予選、2004～2022 年度）

	ベースライン（標準化された評点）							（素点）
		10 人以上	9 人以下	G 級以上	F 級以下			
		(a)	(b)		(a)-(b)	(c)	(d)	
1 番目	−0.117*** (0.005)	−0.106*** (0.006)	−0.236*** (0.028)	0.130*** (0.000)	−0.336*** (0.045)	−0.114*** (0.005)	−0.222*** (0.000)	−0.037*** (0.001)
2 番目	−0.077*** (0.005)	−0.074*** (0.005)	−0.189*** (0.028)	0.115*** (0.000)	−0.370*** (0.046)	−0.073*** (0.005)	−0.297*** (0.000)	−0.025*** (0.001)
3 番目	−0.063*** (0.005)	−0.064*** (0.005)	−0.151*** (0.028)	0.087*** (0.000)	−0.338*** (0.044)	−0.059*** (0.005)	−0.279*** (0.000)	−0.020*** (0.001)
4 番目	−0.046*** (0.005)	−0.049*** (0.005)	−0.113*** (0.027)	0.064** (0.020)	−0.178*** (0.047)	−0.044*** (0.005)	−0.134** (0.005)	−0.014*** (0.001)
5 番目	−0.035*** (0.005)	−0.044*** (0.005)	−0.101*** (0.028)	0.057** (0.045)	−0.079 (0.052)	−0.034*** (0.005)	−0.045 (0.389)	−0.011*** (0.001)
6 番目	−0.023*** (0.005)	−0.025*** (0.005)	−0.088*** (0.028)	0.063** (0.027)	−0.113** (0.047)	−0.022*** (0.005)	−0.091* (0.054)	−0.005*** (0.002)
7 番目	−0.016*** (0.005)	−0.015*** (0.005)	−0.078*** (0.029)	0.063** (0.032)	−0.076 (0.048)	−0.014*** (0.005)	−0.062 (0.199)	−0.004*** (0.002)
8 番目	−0.011** (0.005)	−0.013** (0.005)	−0.042 (0.031)	0.029 (0.360)	0.031 (0.051)	−0.010** (0.005)	0.041 (0.424)	−0.003* (0.002)
9 番目	−0.008 (0.005)	−0.012** (0.005)			−0.032 (0.053)	−0.008 (0.005)	−0.024 (0.652)	−0.002 (0.002)
10 番目	−0.006 (0.005)				0.040 (0.061)	−0.005 (0.005)	0.045 (0.462)	−0.001 (0.002)
最終 演奏	0.000 (0.000)	0.000 (0.000)	-0.011*** (0.002)		0.001 (0.001)	0.000 (0.000)		0.000 (0.000)
N	1,850,992	1,640,325	210,667		29,023	1,821,969		1,851,017
Adj. R²	0.372	0.390	0.336		0.502	0.373		0.443
F 値	77.4	57.66	50.77		13.42	63.49		68.24

- （注）1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティション予選に参加した演奏者のデータを用いた。ベースライン（全大会・リハ除く演奏順）における演奏順は、欠席者・リハーサルを除く、演奏者の大会・級別の演奏順。評点は大会・級ごとに標準化した。
2. すべての推定において、演奏者、審査員、大会開催年・地区、級、累計エントリー回数（リハーサルを含む）の固定効果、リハーサルをのぞく大会・級別の演奏者数、リハーサルがある大会・級ダミーを制御した。（ ）内は、年・大会開催地区クラスター頑健標準誤差。F 値は固定効果を除く説明変数の同時有意性を年・大会開催地区クラスター頑健 Wald 検定で評価した。
3. t 値は各演奏順の係数の差の 2 標本 Wald 検定。（ ）内は p 値。
4. *p<0.10, **p<0.05, ***p<0.01 をあらわす。

表 5 審査員の経験による異質性（2004～2022 年度）

	ベースライン					
		指導者としての 経験年数	指導者賞受賞 回数	ピアノ指導者 ライセンス (ライセンスが あれば1)	ピアノセミナ ー講師（講師 であれば1)	プロピアニス ト（プロピアニ ストであれば 1)
演奏順 (演奏順が 1～3 番目であれば 1)	-0.073*** (0.003)	-0.069*** (0.006)	-0.076*** (0.003)	-0.078*** (0.009)	-0.074*** (0.003)	-0.073*** (0.003)
主効果		0.004*** (0.000)	0.006*** (0.000)	0.107*** (0.012)	0.107*** (0.012)	-0.042*** (0.015)
交差項		-0.000 (0.000)	0.001* (0.000)	0.006 (0.010)	0.006 (0.010)	0.007 (0.013)
N	1,850,992	1,799,267	1,850,992	1,835,094	1,835,094	1,835,094
Adj. R ²	0.332	0.334	0.333	0.333	0.333	0.332
F 値	138.56	219.85	225.07	225.23	225.23	225.68

(注) 1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティション予選に参加した演奏者のデータを用いた。ベースライン（全大会・リハ除く演奏順）における演奏順は、欠席者・リハーサルを除く、演奏者の大会・級別の演奏順。全大会・リハ含む演奏順は、欠席者を除く、リハーサルを含むすべての演奏者の大会・級別の演奏順。評点は大会・級ごとに標準化した。

2. すべての推定において、演奏者、審査員、大会開催年・地区、級、累計エントリー回数（リハーサルを含む）の固定効果、リハーサルをのぞく大会・級別の演奏者数、リハーサルがある大会・級ダミーを制御した。（ ）内は、年・大会開催地区クラスター頑健標準誤差。F 値は固定効果を除く説明変数の同時有意性を年・大会開催地区クラスター頑健 Wald 検定で評価した。

3. 指導者賞は（ア）地区本選（第二次予選）に 8 名以上の生徒が参加していること、（イ）地区本選（第二次予選）に 6 名以上の生徒が参加し、かつ、そのうち 2 名以上が本選優秀賞以上を受賞していること、（ウ）3 名以上の生徒が本選優秀賞以上を受賞していること、（エ）ソロ部門 G 級、Jr.G 級、Pre 特級、特級で 2 名以上の生徒が全国大会または特級セミファイナルに参加していること、のいずれかに該当する指導者に与えられる賞であり、指導者賞受賞回数はこれを受賞した回数、ピアノ指導者ライセンスは、ピアノ指導者またはピアノ指導者を目指しているものが受検する検定に合格した者、ピアノセミナー講師はピティナが主催、共催、後援するセミナーで講師を務めた経験がある者、を指す。

4. *p<0.10, **p<0.05, ***p<0.01 をあらわす。

表6 リハーサルの有無による異質性（2004～2022年度）

	ベースライン リハ除く演奏順	（全大会・リハ除く演奏順）			
		全大会 リハ含む演奏順	リハなし大会 のみ	リハあり大会 のみ リハ含む演奏順	リハあり大会 のみ リハ除く演奏順
演奏順 （演奏順が1～3番目であれば1）	-0.073*** (0.003)	-0.081*** (0.003)	-0.078*** (0.004)	-0.087*** (0.006)	-0.072*** (0.005)
N	1,850,992	1,941,965	984,493	957,472	866,504
Adj. R ²	0.372	0.378	0.372	0.425	0.416
F 値	256.64	130.16	175.6	111.63	119.86

- （注）1. 2004～2022年度（コロナウィルス感染症対策で中止となった2020年度を除く）までにピティナ・ピアノコンペティションに参加した演奏者と審査員のデータを用いた。ベースライン（全大会・リハ除く演奏順）における演奏順は、欠席者・リハーサルを除く、演奏者の大会・級別の演奏順。全大会・リハ含む演奏順は、欠席者を除く、リハーサルを含むすべての演奏者の大会・級別の演奏順。評点は大会・級ごとに標準化した。
2. すべての推定において、演奏者、審査員、大会開催年・地区、級、累計エントリー回数（リハーサルを含む）の固定効果、リハーサルをのぞく大会・級別の演奏者数、リハーサルがある大会・級ダミーを制御した。（ ）内は、年・大会開催地区クラスター頑健標準誤差。F値は固定効果を除く説明変数の同時有意性を年・大会開催地区クラスター頑健Wald検定で評価した。
3. リハなし大会のみでは、リハーサルがない大会に絞って、欠席者を除くすべての演奏者の大会・級別の演奏順。リハあり大会のみ・リハ含む演奏順は、リハーサルがある大会に絞って、欠席者を除く、リハーサルを含むすべての演奏者の大会・級別の演奏順。リハあり大会のみ・リハ除く演奏順は、リハーサルがある大会に絞って、欠席者を除く、リハーサルを除くすべての演奏者の大会・級別の演奏順。
4. ベースラインと比較した時に係数に統計的に有意な差があるかどうかをt値は各演奏順の係数の差の2標本Wald検定で確認したところ、リハあり大会のみ（リハ含む演奏順）との係数だけが5%水準で統計的に有意な差があった。リハーサルを実施した大会では、演奏順1～3番目に入ることによる順序効果がベースラインよりも大きくなっている。
4. *p<0.10, **p<0.05, ***p<0.01をあらわす。

表7 割り付けの結果

開催日	介入群	対照群	大会数
6月3・4日		14	14
6月9・10・11日	31		31
6月16・17・18日	48		48
6月24・25日		43	43
7月1・2日		37	37
7月8・9日	31		31
7月15・16・17日		27	27
合計	110	121	231

(注) 5月1日時点で開催日が未定の2大会、5月下旬に開催された10大会、8月に開催された3大会を除く231大会（級・大会別では1,470）を割り付けの対象にした。他の大会は土日及び祝祭日に行われているのに対し、5月に実施された予選の10大会は平日に開催されていることから、実験の対象から外すこととした。また8月に開催された3大会はソロではなくデュオの大会のため、実験の対象から外した。

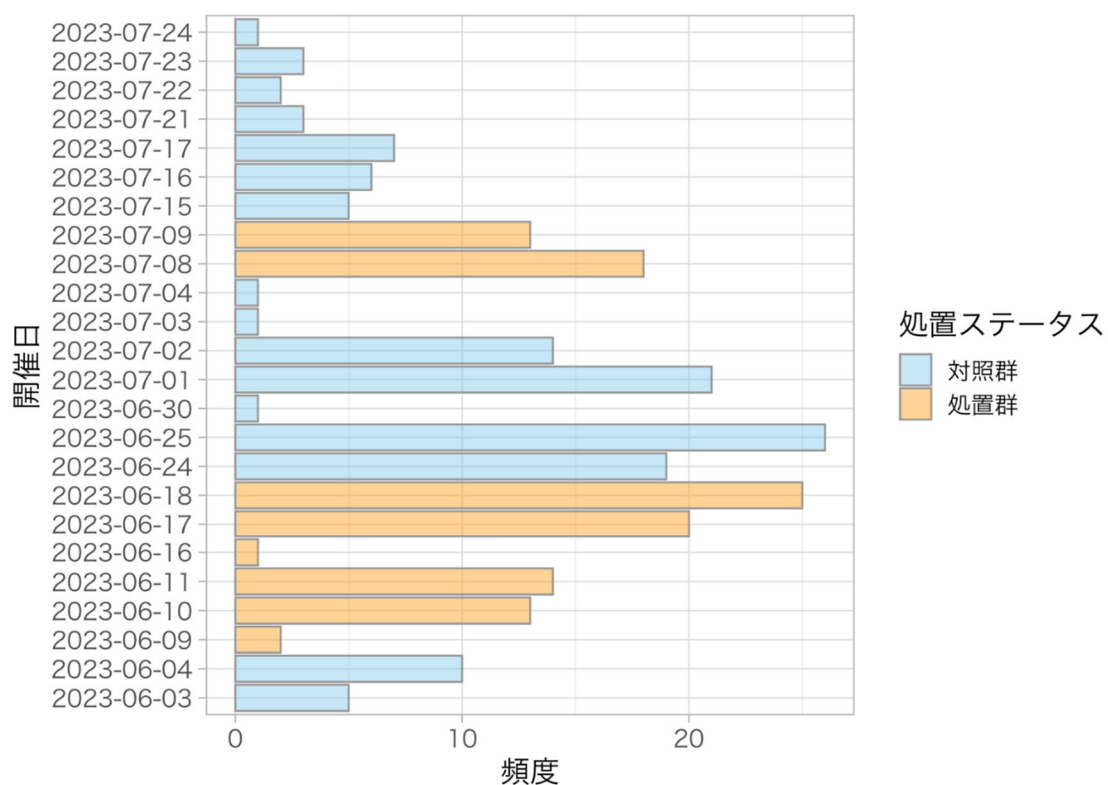


図4 開催日別の大会数

表 8 バランステスト (2023 年度)

		Treatment Group (a)			Control Group (b)			t-value (a)-(b)
		N	Mean	SD	N	Mean	SD	
大会	演奏者数 (大会・級別)	106	18.900	13.800	125	19.600	14.600	0.407
	演奏者数 (大会・級別、リハ除く)	104	18.600	13.800	121	19.700	14.700	0.566
演奏者	年齢	10,587	10.100	3.140	12,836	10.300	3.440	-2.822***
	性別	10,588	0.759	0.427	12,836	0.760	0.427	0.066
	大会エントリー回数	10,589	1.260	0.458	12,838	1.380	0.498	-19.168***
	リハーサルエントリー回数	10,589	0.022	0.157	12,838	0.040	0.231	-6.867***
	初出場	10,589	0.371	0.483	12,838	0.327	0.469	7.026***
指導者	指導者としての経験年数	529	22.600	10.500	628	22.300	10.700	0.449
	女性	532	0.786	0.411	629	0.769	0.422	0.663
	指導者賞受賞回数	532	5.990	8.480	629	6.120	8.360	-0.255
	ピアノ指導者ライセンス	532	0.803	0.398	629	0.798	0.402	0.193
	ピアノセミナー講師	532	0.143	0.350	629	0.132	0.339	0.536
	プロピアニスト	532	0.047	0.212	629	0.007	0.250	-1.460

(注) 1. 2023 にピティナ・ピアノコンペティションに参加した演奏者と審査員のデータを用いた。変数の定義は表 1 と同じ。

2. 指導者賞は (ア) 地区本選 (第二次予選) に 8 名以上の生徒が参加していること、(イ) 地区本選 (第二次予選) に 6 名以上の生徒が参加し、かつ、そのうち 2 名以上が本選優秀賞以上を受賞していること、(ウ) 3 名以上の生徒が本選優秀賞以上を受賞していること、(エ) ソロ部門 G 級、Jr.G 級、Pre 特級、特級で 2 名以上の生徒が全国大会または特級セミファイナルに参加していること、のいずれかに該当する指導者に与えられる賞であり、指導者賞受賞回数はこれを受賞した回数、ピアノ指導者ライセンスは、ピアノ指導者またはピアノ指導者を目指しているものが受検する検定に合格した者、ピアノセミナー講師はピティナが主催、共催、後援するセミナーで講師を務めた経験がある者、を指す。

3. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

表9 介入の効果

分析の対象	ベースライン リハ除く演奏順		全大会 リハ含む演奏順	リハなし大会 のみ	リハあり大会 のみ リハ含む演奏順	リハあり大会 のみ リハ除く演奏順	(参考：予選通過) ベースライン	(参考：予選奨励賞) ベースライン
演奏順D (演奏順 1～3 番目=1)	-0.042*** (0.016)	-0.053*** (0.018)	-0.047*** (0.018)	-0.042** (0.020)	-0.094* (0.048)	-0.101** (0.041)	0.005 (0.012)	0.025* (0.014)
処置群D (処置群=1)	0.009** (0.005)	0.060*** (0.011)	0.060*** (0.011)	0.043*** (0.013)	0.100*** (0.032)	0.097*** (0.033)	0.032*** (0.005)	0.015** (0.007)
演奏順D×処置群D	-0.035 (0.024)	-0.029 (0.024)	-0.024 (0.024)	-0.028 (0.026)	0.020 (0.070)	-0.017 (0.063)	-0.007 (0.017)	0.005 (0.021)
リハーサル (リハあり=1)			✓					
コントロール変数		✓	✓	✓	✓	✓	✓	✓
固定効果		✓	✓	✓	✓	✓	✓	✓
N (演奏回数×スコア) Adj. R ²	113,149 0.001	113,134 0.119	114,842 0.119	89,127 0.119	25,715 0.136	24,007 0.136	113,142 0.073	67,564 0.023

(注) 1. ベースライン (全大会・リハ除く演奏順) における演奏順は、欠席者・リハーサルを除く、演奏者の大会・級別の演奏順。全大会・リハ含む演奏順は、欠席者を除く、リハーサルを含むすべての演奏者の大会・級別の演奏順。

2. コントロール変数は、エントリー回数、リハーサルエントリー回数、表8で2群の間に統計的に有意な差がある変数に加え、大会規模 (欠席・リハ除く最終演奏者の演奏順)。固定効果は、級、審査員、大会開催地区、累計エントリー回数の固定効果。予選通過は予選を通過して本選に進んだかを表すダミー変数、予選奨励賞は予選不通過者のうち上位に与えられる賞を受賞したかをあらわすダミー変数。

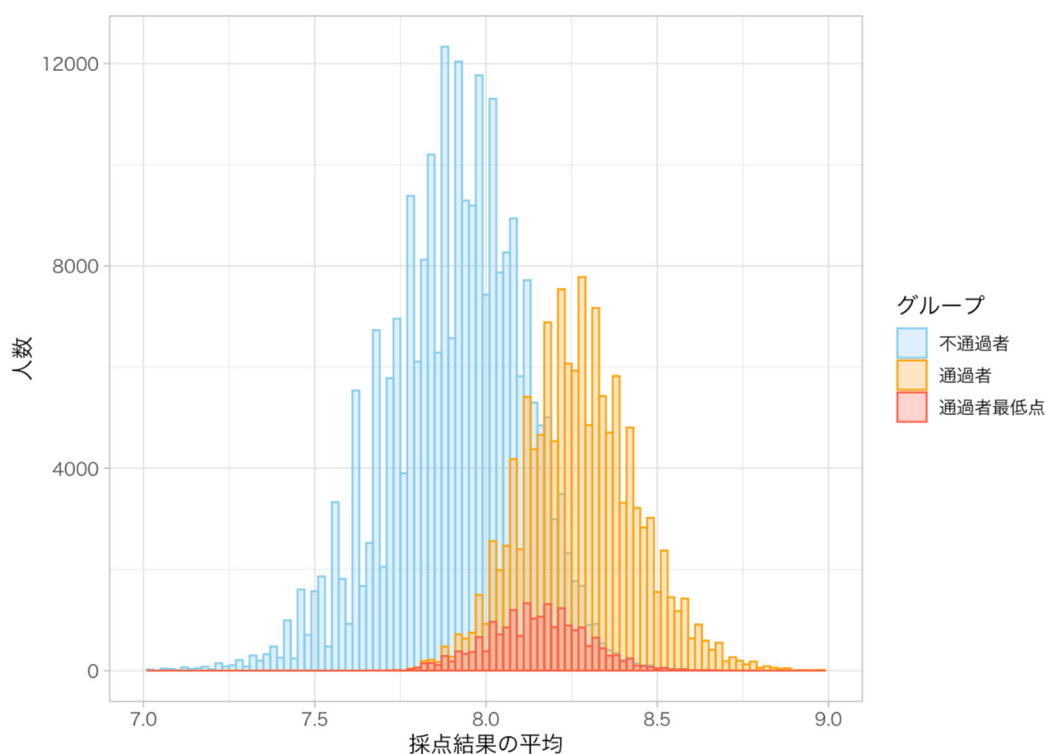
3. 大会ごとに標準誤差をクラスタリングした。

4. *p<0.10, **p<0.05, ***p<0.01 をあらわす。

表 10 介入の異質性（大会の属性）

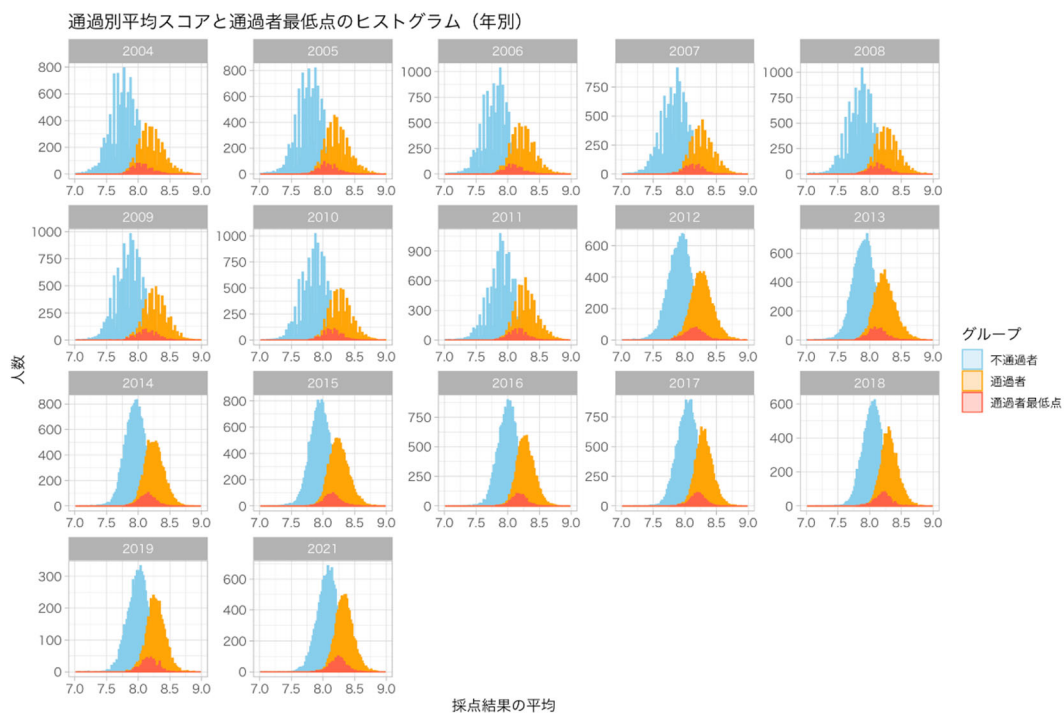
	2 回以上 全大会 リハ除く演奏順	1 グループ目 全大会 リハ除く演奏順	9 人以下 全大会 リハ除く演奏順	G 級以上 全大会 リハ除く演奏順	G 級以上のみ リハ除く演奏順	F 級以下のみ リハ除く演奏順
演奏順 (演奏順が 1～3 番目であれば 1)	-0.049*** (0.018)	-0.041** (0.020)	-0.049** (0.023)	-0.035** (0.017)	-0.535*** (0.091)	-0.044** (0.018)
treatment (介入群であれば 1)	0.080*** (0.011)	0.073*** (0.012)	0.071*** (0.011)	-0.062*** (0.011)	-0.325* (0.155)	0.065*** (0.011)
演奏順×treatment	-0.033 (0.024)	-0.048* (0.027)	-0.051 (0.032)	-0.035 (0.024)	0.267 (0.156)	-0.034 (0.024)
演奏順×2 回以上の介入	-0.038 (0.070)					
演奏順×treatment×大会開始 1 グループ目の介入		0.054 (0.055)				
演奏順×treatment×9 人以下			0.064 (0.059)			
演奏順×treatment×G 級以上				0.271** (0.125)		
N (演奏回数×スコア)	113,134	113,134	113,134	113,134	2,643	110,491
Adj. R ²	0.105	0.104	0.104	0.118	0.153	-0.120

- (注) 1. ベースライン（全大会・リハ除く演奏順）における演奏順は、欠席者・リハーサルを除く、演奏者の大会・級別の演奏順。全大会・リハ含む演奏順は、欠席者を除く、リハーサルを含むすべての演奏者の大会・級別の演奏順。
2. 表 9 の 2 列目の推定と同じ定式化。交差項を含む回帰式では、対応する主効果も含めて推定している。
3. 大会ごとに標準誤差をクラスタリングした。
4. *p<0.10, **p<0.05, ***p<0.01 をあらわす。



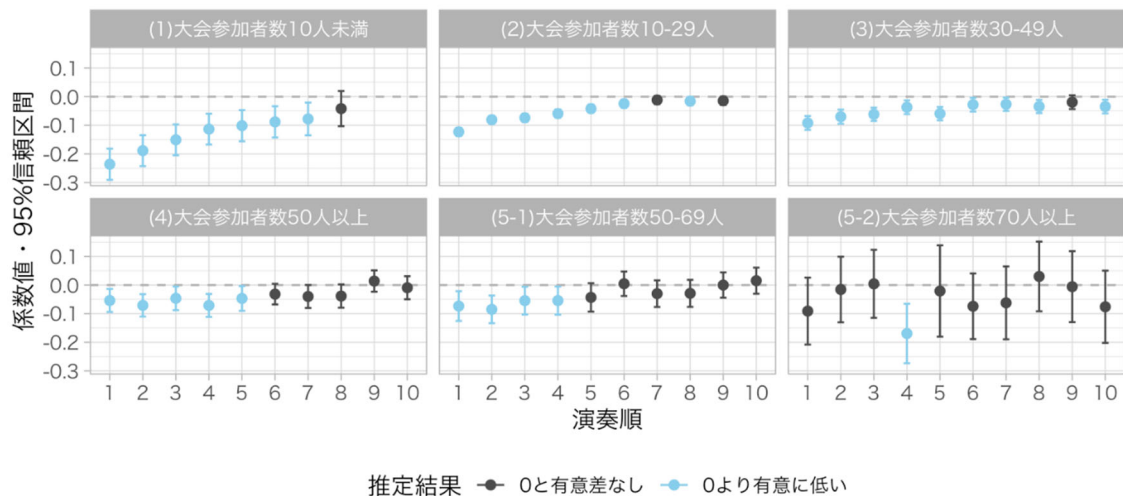
Appendix 1 審査員の平均点の分布（全体）

- (注) 1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティションに参加したのべ約 191 万人の演奏者のデータを用いた。審査員による各演奏者の評点の素点の分布をあらわす。
2. 青の分布は不合格者。黄色の分布は本選に進むことができる合格者。赤は合格最低点の分布である。

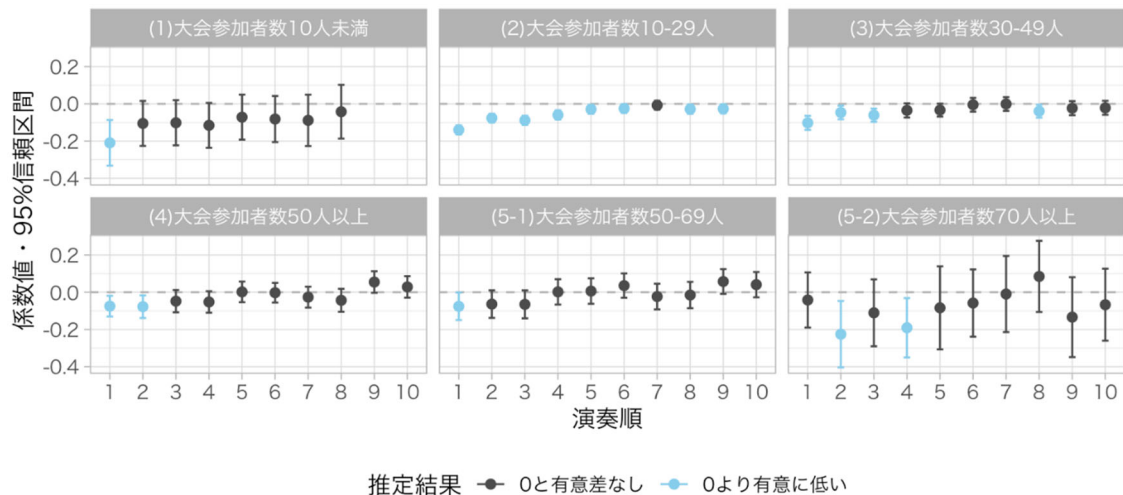


Appendix 1（参考） 審査員の平均点の分布（年別）

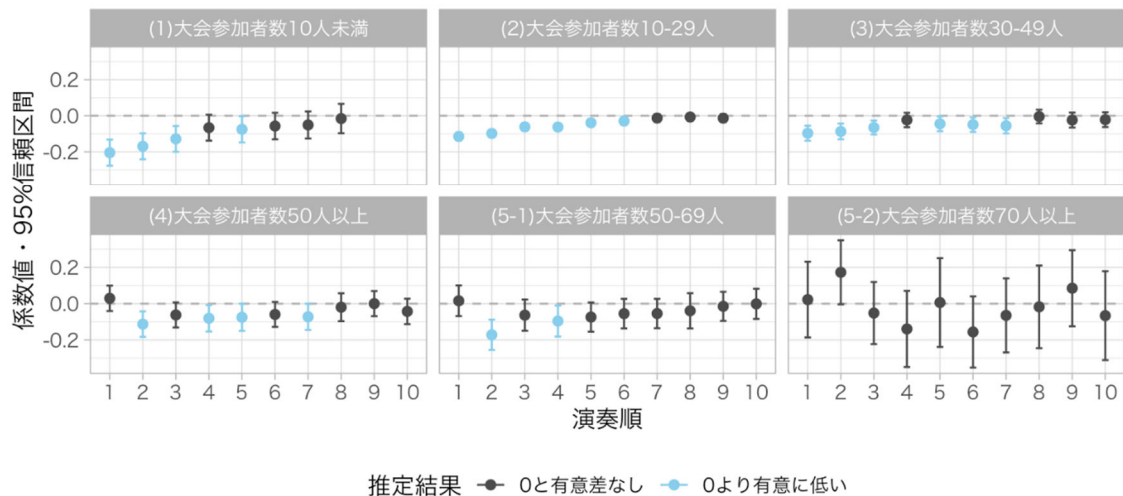
ベースライン：全大会（リハ含む）のリハ除く演奏順：大会参加者数による比較



リハーサルのある大会のリハ除く演奏順：大会参加者数による比較

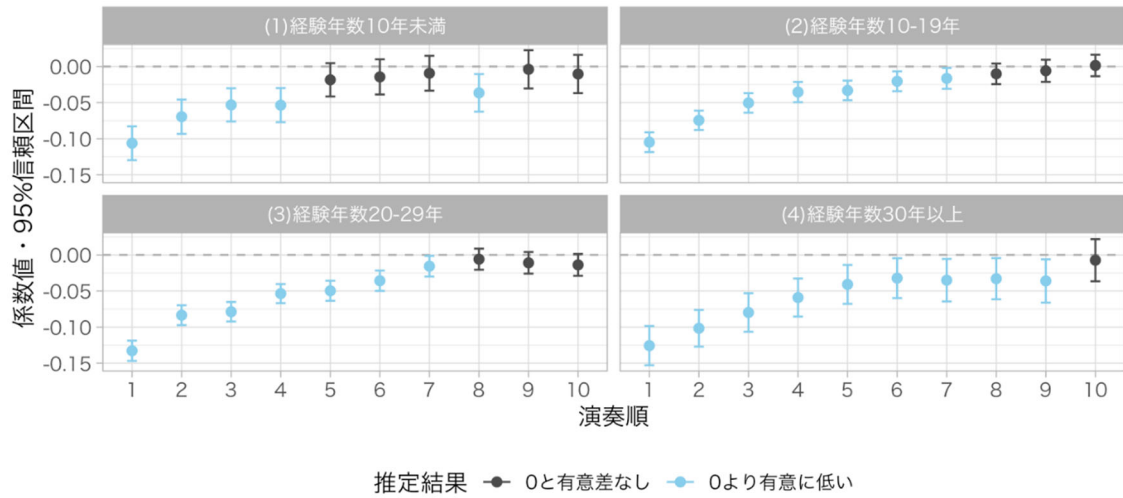


リハーサルのない大会：大会参加者数による比較

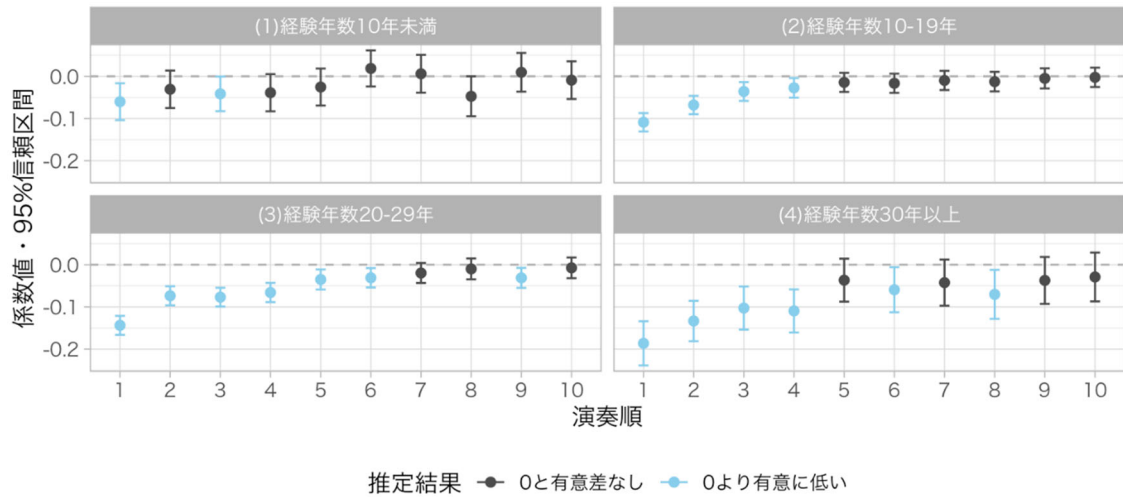


Appendix 2 大会参加者数の異質性

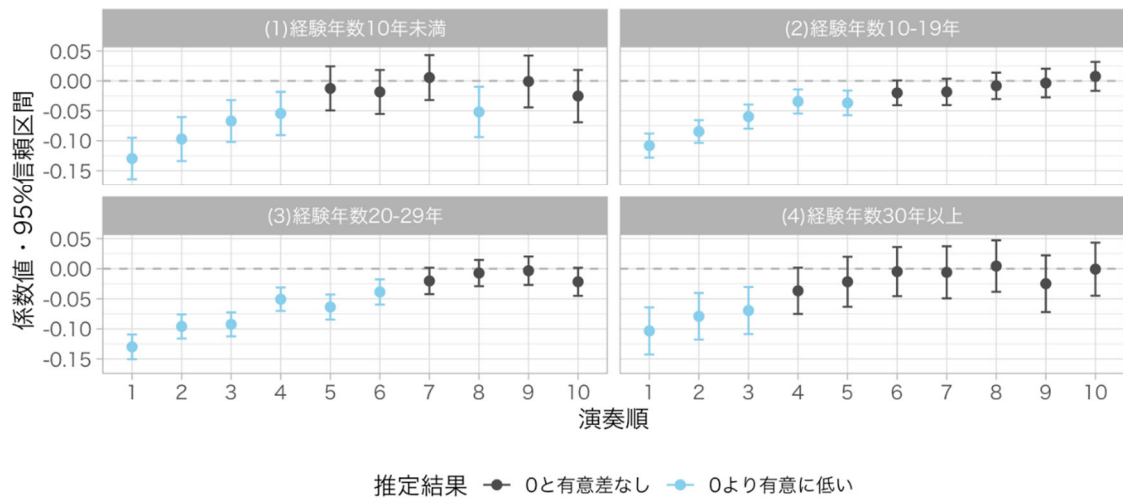
ベースライン：全大会（リハ含む）のリハ除く演奏順：経験年数による比較



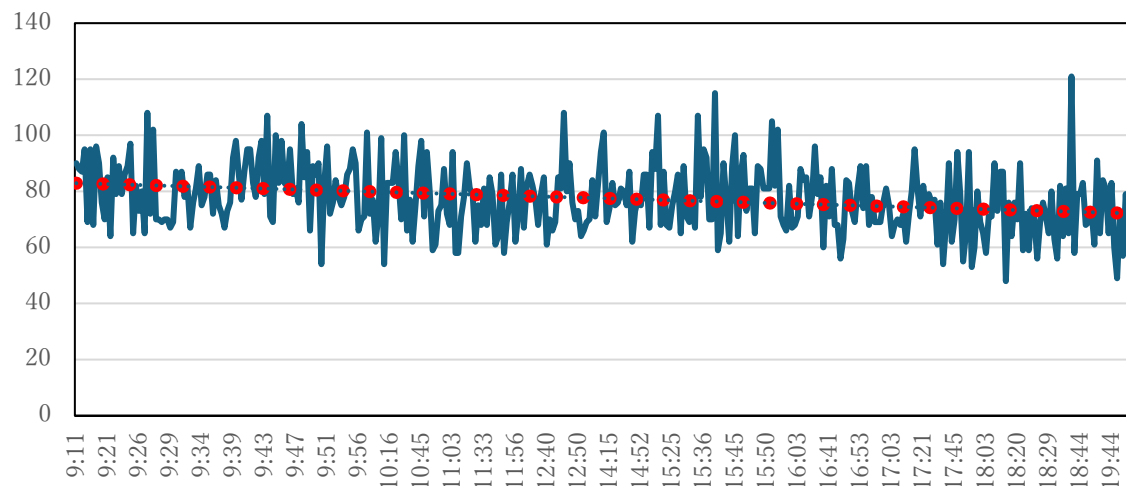
リハーサルのある大会のリハ除く演奏順：経験年数による比較



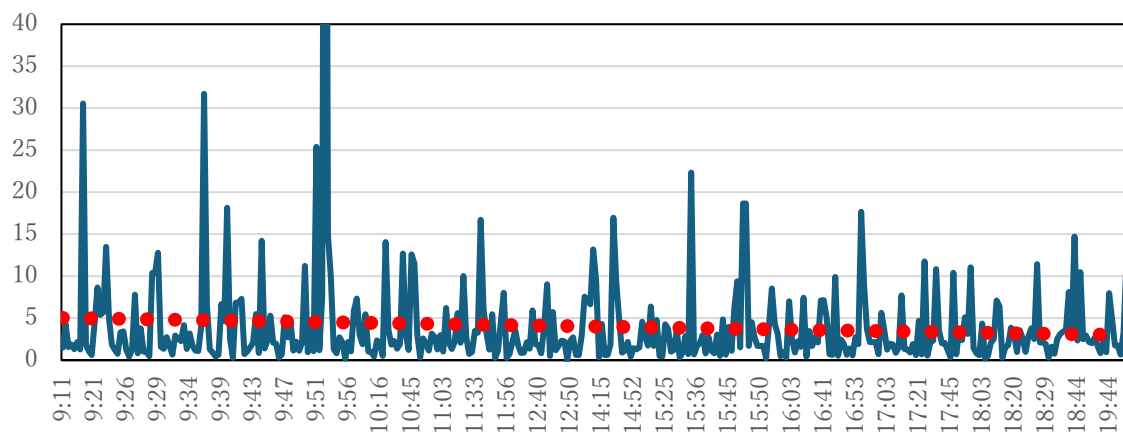
リハーサルのない大会：経験年数による比較



Appendix 3 審査員経験年数別の異質性



各審査時間帯の平均脈拍数



各審査時間帯における平均低周波成分（LF）/高周波成分（HF）

- (注) 1. 日立システムズ株式会社が疲労科学研究所と共同開発した自律神経測定器を用いて、指先から心電波・脈波を計測したもの。両手人差し指の指先をセンサに振れる形で計測でき、日内変動を捉えることができる。自律神経の病気で服薬している人や、不整脈のある人は正確に計測できないことから、計測の対象外とした。
2. 心拍変動を周波数解析し、低周波成分（LF）で交感神経、高周波成分（HF）で副交感神経を把握し、交感神経と副交感神経のバランス値で疲労を計測している。心拍数も LF/HF 比率も、数値が高くなれば疲労度やストレスが増すことを意味する。
3. 全国で予選の後に行われた本選において、106 名の審査員に、測定器を用いて、午前と午後に少なくとも 1 回の測定を依頼した（約 370 回の計測が行われた）。上記のグラフは、地域、開催日、審査員の固定効果などは制御せず、時間の経過とともに脈拍数や LF/HF の推移をみたものの。赤線は直線近似。

Appendix 4 審査員の疲労度の測定

Appendix 5 コンクールの時間経過と審査員の疲労度

	心拍数	LF/HF
9 時からの経過時間（分）	-0.013*** (0.002)	-0.001 (0.001)
地域固定効果	✓	✓
日付固定効果	✓	✓
審査員固定効果	✓	✓
N	370	370
Adj. R ²	0.810	0.383

(注) 1. 審査員ごとに標準誤差をクラスタリングした。

2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ をあらわす。

Appendix 6 割り当てメカニズムの統計的検証

	審査員個人		審査員 5 人の平均（演奏者単位）	
	標準化評点	合格 (合格であれば=1)	標準化評点	合格 (合格であれば=1)
直前演奏者の標準化得点	0.166*** (0.003)		0.029* (0.014)	
直前演奏者の合格		0.012*** (0.001)		0.010 (0.009)
直前 5 回の肯定的な決定の割合	✓	✓	✓	✓
指導者としての経験年数	✓	✓	✓	✓
女性	✓	✓	✓	✓
指導者賞受賞回数	✓	✓	✓	✓
ピアノ指導者ライセンス	✓	✓	✓	✓
ピアノセミナー講師	✓	✓	✓	✓
プロピアニスト	✓	✓	✓	✓
N	1,686,228	1,686,228	338,002	331,206
Adj. R ²	0.074	0.152	0.181	0.061

(注) 1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティションに参加した演奏者と審査員のデータを用いた。評点は大会・級ごとに標準化した。リハーサルは分析から除いた。標準誤差は審査員クラスター頑健標準誤差を用いた。大会および演奏者の固定効果を制御した。

2. 審査員の属性の定義については、表 5 と同様。

3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ をあらわす。

Appedix 7 演奏者が 1 人の大会・級と 2 人以上の大会・級の比較 (2004～2022 年度)

		演奏者が 1 人の大会・級 (a)			2 人以上の大会・級 (b)			t-value (a)-(b)
		N	Mean	SD	N	Mean	SD	
大会	リハーサルの有無	331	0.199	0.400	16246	0.374	0.484	-7.845***
演奏者	年齢	481	14.900	3.600	20270	11.300	3.780	21.967***
	性別	481	0.848	0.359	20255	0.830	0.376	1.101
	大会エントリー回数	481	1.230	0.423	20270	1.260	0.440	-1.526
	リハーサルエントリー回数	481	0.015	0.164	20270	0.022	0.156	-0.921
	初出場	481	0.208	0.406	0.282	0.450	20270	-3.935***
指導者	指導者としての経験年数	325	18.70	8.640	15730	19.000	8.590	-0.542
	女性	329	0.796	0.403	16096	0.805	0.396	-0.382
	指導者賞受賞回数	331	4.930	7.170	16246	4.560	6.790	0.937
	ピアノ指導者ライセンス	329	0.921	0.270	16096	0.927	0.261	-0.376
	ピアノセミナー講師	329	0.146	0.354	16096	0.161	0.368	-0.784
	プロピアニスト	329	0.0547	0.228	16096	0.045	0.208	0.749

(注) 1. 2004～2022 年度 (コロナウィルス感染症対策で中止となった 2020 年度を除く) までにピティナ・ピアノコンペティションに参加した演奏者と審査員のデータを用いた。変数の定義は表 1 と同じ。

2. 指導者賞は (1) 予選を通過し本選に進んだ生徒が 8 名以上、(2) 本選に 6 名以上の生徒が参加し、かつ、そのうち 2 名以上が本選優秀賞以上を受賞のいずれかの指導者に与えられる賞であり、指導者賞受賞回数はこれを受賞した回数、ピアノ指導者ライセンスは、ピアノ指導者またはピアノ指導者を目指しているものが受検する検定に合格した者、ピアノセミナー講師はピティナが主催、共催、後援するセミナーで講師を務めた経験がある者、を指す。

3. * $p<0.1$; ** $p<0.05$; *** $p<0.01$

Appendix 8 演奏者が1人の大会・級と2人以上の大会・級の順序効果の比較（2004～2022年度）

	標準化スコア				予選通過率			
演奏順1番目	0.109*** (0.013)	0.114*** (0.016)	0.109*** (0.017)	0.107*** (0.018)	0.481*** (0.016)	0.343*** (0.017)	0.341*** (0.017)	0.340*** (0.017)
大会開催年・地区・級 FE		✓	✓	✓		✓	✓	✓
リハーサル（リハあり=1）			✓	✓			✓	✓
演奏者数			✓	✓			✓	✓
演奏者の年齢			✓	✓			✓	✓
その他の共変量				✓				✓
N	103,336	103,336	103,336	103,336	103,361	103,361	103,361	103,361
Adj. R ²	0.000	0.006	0.037	0.047	0.022	0.075	0.107	0.118
F 値	68.64	49.76	354.33	355.47	927.56	405.51	322.07	306.46

（注）1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2021 年度を除く）までにピティナ・コンペティションに参加した演奏者と審査員のデータを用いた。

2. 演奏者1名のみでの大会は欠席者・リハーサルを除き、演奏者が1名の場合と定義した。

3. 標準誤差の計算には大会クラスター頑健標準誤差を用いた。F 値は固定効果を除く説明変数の同時有意性を大会クラスター頑健な Wald 検定で評価した。

Appendix 9 順序効果（本選、2004～2022 年度）

	ベースライン（標準化された評点）							（素点）
		10 人以上	9 人以下		G 級以上	F 級以下		
		(a)	(b)	(a)-(b)	(c)	(d)	(c)-(d)	
1 番目	−0.072*** (0.011)	−0.070*** (0.011)	−0.441** (0.186)	0.371** (0.046)	−0.444*** (0.100)	−0.069*** (0.011)	−0.375*** (0.000)	−0.023*** (0.003)
2 番目	−0.112*** (0.011)	−0.109*** (0.011)	−0.421** (0.175)	0.312* (0.075)	−0.761*** (0.101)	−0.108*** (0.011)	−0.653*** (0.000)	−0.033*** (0.003)
3 番目	−0.097*** (0.011)	−0.092*** (0.011)	−0.408* (0.243)	0.316 (0.194)	−0.401*** (0.118)	−0.095*** (0.011)	−0.306** (0.010)	−0.031*** (0.003)
4 番目	−0.075*** (0.010)	−0.072*** (0.010)	−0.147 (0.185)	0.075 (0.686)	−0.196* (0.101)	−0.074*** (0.010)	−0.122 (0.225)	−0.024*** (0.003)
5 番目	−0.079*** (0.010)	−0.075*** (0.011)	−0.135 (0.195)	0.060 (0.759)	−0.362** (0.154)	−0.078*** (0.010)	−0.284* (0.066)	−0.024*** (0.003)
6 番目	−0.074*** (0.011)	−0.072*** (0.011)	−0.115 (0.226)	0.043 (0.849)	−0.158 (0.136)	−0.073*** (0.011)	−0.085 (0.535)	−0.022*** (0.003)
7 番目	−0.060*** (0.011)	−0.058*** (0.011)	−0.130 (0.173)	0.072 (0.678)	−0.145 (0.121)	−0.060*** (0.011)	−0.085 (0.482)	−0.016*** (0.003)
8 番目	−0.048*** (0.011)	−0.047*** (0.011)	−0.513** (0.215)	0.466** (0.030)	0.038 (0.090)	−0.048*** (0.011)	0.086 (0.342)	−0.013*** (0.003)
9 番目	−0.066*** (0.010)	−0.065*** (0.010)			−0.120 (0.146)	−0.066*** (0.010)	−0.054 (0.712)	−0.020*** (0.003)
10 番目	−0.048*** (0.011)				0.010 (0.078)	−0.049*** (0.011)	0.059 (0.451)	−0.012*** (0.003)
最終 演奏	0.002*** (0.000)	0.002*** (0.000)	0.005 (0.043)		0.001 (0.008)	0.002*** (0.000)		−0.000 (0.000)
N	1,574,758	1,566,810	7,946		12,916	1,561,842		1,574,758
Adj. R ²	0.339	0.339	0.507		0.496	0.340		0.385
F 値	27.14	28.86	5.09		8.37	25.98		28.74

- （注）1. 2004～2022 年度（コロナウィルス感染症対策で中止となった 2020 年度を除く）までにピティナ・ピアノコンペティション本選に参加した演奏者のデータを用いた。ベースラインにおける演奏順は、欠席者を除く、演奏者の大会・級別の演奏順。評点は大会・級ごとに標準化した。なお、本選では級によらず、リハーサルは行われない。
2. すべての推定において、演奏者、大会開催年・地区、級、審査員、累計エントリー回数（本大）の固定効果、大会規模（級別）、を制御した。（ ）内は、年・大会開催地区クラスター頑健標準誤差。 F 値は固定効果を除く説明変数の同時有意性を年・大会開催地区クラスター頑健 Wald 検定で評価した。
3. t 値は各演奏順の係数の差の 2 標本 Wald 検定。（ ）内は p 値。
4. *p<0.10, **p<0.05, ***p<0.01 をあらわす。

Appendix 10 介入の異質性（指導者の属性）

	指導者としての 経験年数	女性	指導者賞受賞回 数	ピアノ指導者ラ イセンス	ピアノセミナー 講師	プロピアニス ト
演奏順 (演奏順が 1～3 番目であれば 1)	－0.088*** (0.032)	－0.053* (0.030)	－0.082*** (0.020)	－0.106*** (0.032)	－0.058*** (0.018)	－0.059*** (0.018)
treatment (介入群であれば 1)	－0.004 (0.028)	－0.008 (0.029)	0.058*** (0.013)	0.040 (0.028)	0.047*** (0.009)	0.036*** (0.008)
演奏順×treatment	0.011 (0.043)	－0.019 (0.043)	－0.000 (0.026)	0.041 (0.043)	－0.016 (0.024)	－0.017 (0.024)
演奏順×treatment×指導者として の経験年数	－0.001 (0.002)					
演奏順×treatment×女性		0.000 (0.046)				
演奏順×treatment×指導者賞受賞 回数			－0.003 (0.002)			
演奏順×treatment×ピアノ指導者 ライセンス				－0.072 (0.046)		
演奏順×treatment×ピアノセミナ ー講師					－0.018 (0.053)	
演奏順×treatment×プロピアニス ト						－0.041 (0.074)
N（演奏回数×スコア） Adj. R ²	110,515 0.047	110,627 0.046	11,0627 0.046	110,627 0.047	110,627 0.046	110,627 0.046

- (注) 1. ベースライン（全大会・リハ除く演奏順）における演奏順は、欠席者・リハーサルを除く、演奏者の大会・級別の演奏順。全大会・リハ含む演奏順は、欠席者を除く、リハーサルを含むすべての演奏者の大会・級別の演奏順。
2. 表 9 の 2 列目の推定と同じ specification。交差項を含む回帰式では、対応する主効果も含めて推定している。
3. 大会ごとに標準誤差をクラスタリングした。
4. *p<0.10, **p<0.05, ***p<0.01 をあらわす。

表 10 介入の異質性（大会の属性）

	2 回以上 全大会 リハ除く演奏順	1 グループ目 全大会 リハ除く演奏順	9 人以下 全大会 リハ除く演奏順	G 級以上 全大会 リハ除く演奏順	G 級以上のみ リハ除く演奏順	F 級以下のみ リハ除く演奏順
演奏順 (演奏順が 1～3 番目であれば 1)	-0.049*** (0.018)	-0.041** (0.020)	-0.049** (0.023)	-0.035** (0.017)	-0.535*** (0.091)	-0.044** (0.018)
treatment (介入群であれば 1)	0.080*** (0.011)	0.073*** (0.012)	0.071*** (0.011)	-0.062*** (0.011)	-0.325* (0.155)	0.065*** (0.011)
演奏順×treatment	-0.033 (0.024)	-0.048* (0.027)	-0.051 (0.032)	-0.035 (0.024)	0.267 (0.156)	-0.034 (0.024)
演奏順×2 回以上の介入	-0.038 (0.070)					
演奏順×treatment×大会開始 1 グループ目の介入		0.054 (0.055)				
演奏順×treatment×9 人以下			0.064 (0.059)			
演奏順×treatment×G 級以上				0.271** (0.125)		
N (演奏回数×スコア) Adj. R ²	113,134 0.105	113,134 0.104	113,134 0.104	113,134 0.118	2,643 0.153	110,491 -0.120

- (注) 1. ベースライン（全大会・リハ除く演奏順）における演奏順は、欠席者・リハーサルを除く、演奏者の大会・級別の演奏順。全大会・リハ含む演奏順は、欠席者を除く、リハーサルを含むすべての演奏者の大会・級別の演奏順。
2. 表 9 の 2 列目の推定と同じ specification。交差項を含む回帰式では、対応する主効果も含めて推定している。
3. 大会ごとに標準誤差をクラスタリングした。
4. *p<0.10, **p<0.05, ***p<0.01 をあらわす。

演奏順が前のほうでは、 点数が低くなる傾向があります。



演奏順が1～3番目の奏者と、それ以降の奏者の点数差（予選）



こちらの動画
をご覧ください
(5分)



（注）2004～2019、2021年の17年間の予選の全ての級・大会のデータを使用。

- ◆ 全ての級・大会で演奏順が最初の方の奏者が不利になっています。
- ◆ 演奏順1番がもっとも不利な傾向がありますが、どの大会でも、最初の3人は、それ以降の奏者よりも（統計的に有意に）スコアが低い傾向があります。
- ◆ 演奏順の1～3番目の奏者と、それ以降の奏者の平均点は0.1点以下の差であることがほとんどですが、このわずかな差が予選突破に影響することもあります。例えば、D級では、演奏順の1～3番目の奏者はそれ以降の奏者よりも、平均で0.07点低いのですが、D級の合格最低点と不合格の差もまた0.07点です。本来であれば合格していたはずの奏者が、演奏順が最初の方だったことによって合格できなかったということになります。
- ◆ スコアを提出する前に、もう一度最初の奏者のスコアの見直しをお願いします。

Appendix 11 動画の内容のスク립ト

ピティナ・ピアノコンペティションで審査を行う審査員の皆様、

はじめまして。こんにちは。私たちは応用ミクロ経済学を専門に研究をしている東京大学、佐賀大学、慶應義塾大学の研究者です。今回、ピティナのご協力をいただき、コンペティションの現場において、審査に関する調査を行わせていただくことになりました。ご協力いただけましたら幸いです。

審査の前に、まずはこの5分ほどの動画をご覧ください。この映像は、コンクールを審査してくださる審査員の先生方に対する情報提供の一環としてご覧いただくものです。

ピティナのウェブサイトに掲載された2006年のショパン国際コンクールのレポートには、レポーターの音楽ジャーナリストの方による次のような記述があります。

「第2次予選の結果は、中国人3名、ロシア人3名のコンテストが本選に進出。1日目に弾いた11名の中から2名、2日目に弾いた5名の中から4名というのは、少し不公平な気がした。1日目の全体のレベルは非常に高く、とくに日本人の2人のコンテストはいずれも質の高い演奏をしていただだけに、どちらも本選に残らなかったのは意外であった。演奏順が遅いほど有利というのは、どのコンクールにも言える傾向のようだ。」

「演奏順が遅いほど有利というのは、どのコンクールにも言える傾向」—これは本当でしょうか。これまで指導者や審査員としての経験がおありになる先生方の中にはそうした印象をお持ちの方がいるかもしれません。実際にSNS上でも、「指導者の先生に演奏順が1番だと絶対に1位にはなれないので残念だったね」と言われたというようなエピソードは数多くみられています。

私たちの研究チームは、ピティナと協力して、2004~2019, 2021年の17年間の、全ての級・大会のデータをもとに、演奏順と審査結果の関係について分析を行いました。例えば、今ご覧いただいている図は、B級、C級という2つの級で、演奏順が40番目くらいまでの奏者の点数です。グラフの中の点数は、合格最低点と提出スコアの差が大会ごとに平均0になるようにしたスコアです。演奏順は抽選でランダムに決まりますから、演奏順は点数と無関係のはずですが、グラフを見ていただくとわかるとおり、最初の数人は明らかに点数が低いことがわかります。ここでは、予選のB級、C級の結果しかお見せしていませんが、予選、本選、全国大会の全ての級で同じことが生じています。どうも、1番だけが不利になるというわけではなく、最初の数人が低く、後ろになればなるほど徐々にその不利が解消される傾向があるようです。

実は演奏順がパフォーマンスに影響するのはピアノコンクールだけで見られる現象ではありません。バイオリンなどのほかの音楽コンクールでも同様のことが指摘されていますし、フィギュアスケート、のど自慢、M1グランプリなどでも同じことが指摘されている

のです。海外で行われたポップアイドルのコンテストのデータを用いて行われた学術研究でも、図が示す通り、最初の出場者の点数が低い傾向があり、後半になればなるほど有利になっていることがわかっています。

どうしてこのようなことが生じるのでしょうか。現時点では有力な仮説は2つほどあります。1つは「疲労」です。演奏順が後ろになるほど、審査員が消耗し、寛容な選択をしやすくなるというものです。もう1つは、審査員が一貫性のある決定を行おうとすることによって生じるというものです。事前に得点分布がわからないので、最初は何の程度の成績が満点に該当するかがわかりません。このため、序盤では極端な判断や評価を避け、後半ではその基準を元により良いものに高い得点を与え、一貫性を保とうとするというもので、これらは「コンテスト」の文脈ではよく見られることがわかっています。

ピアノコンクールでどのようなことが起きているかはまだ詳しくわかりません。それが今回、ピティナの協力を得て、私たち大学研究者が研究に乗り出した理由です。音楽やスポーツ、芸術における評価だけでなく、受験や就職の際の面接などでも同じことが生じているかもしれませんから、順番がパフォーマンスに影響を与えるとみられる分野は広範にわたります。今回の研究の成果が社会に与える影響は大きいと考えています。

この動画をご覧いただいた地区では、審査員の皆様に「演奏順が点数に影響している」ということを念頭に置いて審査に臨んでいただく、という試みを行います。最初の数名の奏者について、あとから演奏を振り返るなどして、最初の数名の奏者が不利にならないような意識を持って審査をお願いいたします。またお手元のチラシのデータも参考にしてください。予選の中には特定の級の奏者が数名しかいない地区もありますが、奏者が少ない場合は演奏順の影響は殆どありませんので、通常通り審査をしていただければ大丈夫です。皆様のご協力に感謝申し上げます。