Exploring Potential Schedule-Related and Gender Biases in Ophthalmology Residency Interview Scores

Chih-Chiun J. Chang, MD[1] Omar Moussa, MD[2] Royce W. S. Chen, MD[2] Lora R. Dagi Glass, MD[2] George A. Cioffi, MD[2] Jeffrey M. Liebmann, MD[2] Bryan J. Winn, MD[1,2,3]

[1] Department of Ophthalmology, University of California, San Francisco, San Francisco, California
[2] Department of Ophthalmology, Columbia University Irving Medical Center, New York-Presbyterian Hospital, New York, New York
[3] Ophthalmology Section, Surgical Service, San Francisco Veterans Affairs Medical Center, San Francisco, California

Address for correspondence Bryan J. Winn, MD, Department of Ophthalmology, University of California, San Francisco, 490 Illinois Street, 5th Floor, San Francisco, CA 94143 (e-mail: bryan.winn@ucsf.edu).

J Acad Ophthalmol 2022;14:e153–e165.

Abstract

Purpose Prior studies have revealed grading discrepancies in evaluation of personal statements and letters of recommendation based on candidate's race and gender. Fatigue and the end-of-day phenomenon can negatively impact task performance but have not been studied in the residency selection process. Our primary objective is to determine whether factors related to interview time and day as well as candidate's and interviewer's gender have a significant effect on residency interview scores.

Methods Seven years of ophthalmology residency candidate evaluation scores from 2013 to 2019 were collected at a single academic institution, standardized by interviewer into relative percentiles (0–100 point grading scale), and grouped into the following categories for comparisons: different interview days (Day 1 vs. Day 2), morning versus afternoon (AM vs. PM), interview session (Day 1 AM/PM vs. Day 2 AM/PM), before and after breaks (morning break, lunch break, and afternoon break), residency candidate's gender, and interviewer's gender.

Results Candidates in the morning sessions were found to have higher scores than afternoon sessions (52.75 vs. 49.28, $p < 0.001$). Interview scores in the early morning, late morning, and early afternoon were higher than late afternoon scores (54.47, 53.01, 52.15 vs. 46.74, $p < 0.001$). Across all interview years, there were no differences in scores received before and after morning breaks (51.71 vs. 52.83, $p = 0.49$), lunch breaks (53.01 vs. 52.15, $p = 0.58$), and afternoon breaks (50.35 vs. 48.30, $p = 0.21$). No differences were found in scores received by female versus male candidates (51.55 vs. 50.49, $p = 0.21$) or scores given by female versus male interviewers (51.31 vs. 50.84, $p = 0.58$).

Conclusion Afternoon residency candidate interview scores, especially late afternoon, were significantly lower than morning scores, suggesting the need to further

Keywords
- ophthalmology residency
- residency interview
- resident selection
- interview time
- interview day
- interview score
- interviewer's fatigue

received
June 21, 2021
accepted
November 10, 2021

DOI https://doi.org/10.1055/s-0042-1744272.
ISSN 2475-4757.

Thieme Medical Publishers, Inc., 333 Seventh Avenue, 18th Floor, New York, NY 10001, USA

study the effects of interviewer's fatigue in the residency interview process. The interview day, presence of break times, candidate's gender, and interviewer's gender had no significant effects on interview score.

Medical school grades, standardized test scores, and honor society status were frequently utilized as screening tools for residency candidate evaluation and interview invitations.[1–3] However, many medical schools have adopted a pass/fail grading system for preclinical courses and clinical clerkships[4] and removed the consideration for Alpha Omega Alpha and Gold Humanism Honor Society. In addition, the United States Medical Licensing Examination Step 1 has now become a pass/fail examination. The reduced number of objective measures for candidate competencies has increased the importance of evaluating noncognitive personality traits such as work ethic, maturity, leadership qualities, and interpersonal communication skills[5] through the personal statement, letters of recommendation, and interview evaluations.[6]

Prior studies have revealed grading discrepancies in evaluation of personal statements and letters of recommendation based on race,[7–9] candidate's gender,[10–14] and evaluator's gender.[15–17] It is important to determine whether similar biases exist in the faculty interview process, as the interview is often considered the most important noncognitive assessment for residency selection and ranking.[18–20] Previous studies have looked at several factors that may influence the evaluation of candidates including the use of open-file versus blinded interviews,[21,22] individual faculty characteristics (e.g., gender, number of publications, and years of clinical practice),[16] candidate's gender,[10] and facial appearance.[23,24]

The scheduling of residency interviews is a time-sensitive process that often rewards quick responders with preferred dates and times. Evaluation of the residency match in internal medicine,[25] emergency medicine,[26] and 10 additional specialties showed that there was no correlation between interview date and a successful match.[27] Other external factors including time of day have not been well studied in the interview process. This study will add to the existing literature by evaluating the effects of interview date, time of day, session, and break times on interview scores and determine whether there are any disparities in grading practices based on candidate's and interviewer's gender.

Methods

Interview Score Database

This study was conducted using 7 years of residency interview data obtained from the ophthalmology residency program at the Columbia University Irving Medical Center/New York-Presbyterian Hospital from 2013 to 2019. It was approved by the Columbia University Irving Medical Center Institutional Review Board and was compliant with protection of individually identifiable information; it also adhered to the tenets of the Declaration of Helsinki as amended in 2013. The available residency interview data contained candidate's photographs and names, interview day and session (AM or PM), interview times associated with each candidate–interviewer pair, times of breaks, waitlist status, and numerical raw scores given by interviewers to candidates on a scale of 0 to 30 from 2013 to 2016 and 0 to 10 scale from 2017 to 2019.

Perceived candidate's gender was recorded based on applicant's photographs as male, female, or indeterminable by one of the authors (C.C.). The names of candidates and interviewers as well as photographs were stripped from the dataset and replaced with deidentified numerical identification (ID) codes prior to analysis. A deidentified dataset containing the following information was used for the primary analysis: candidate's ID, candidate waitlist status, candidate's gender, interview year, interview session (Day 1 AM/PM vs. Day 2 AM/PM), interview position (number of people interviewed before the candidate), interview time, interviewer's ID, interviewer's gender, raw interview score, standardized percentile score, and applicant match status.

Interview Day Schedule

The interview days were conducted on Thursday (interview Day 1) and Friday (interview Day 2). Candidates were present for either a morning (AM) session (8 AM–12 PM) or an afternoon (PM) session (1 PM–5 PM). Candidates typically had a series of five to six interviews, each lasting ~8 to 10 minutes with 2 to 5 minutes between the interviews. Each interview was conducted by a panel of two to three faculty or chief resident interviewers. Candidates either had a tour of the eye institute and campus before or after this series of interviews. One 10-minute break was included midway through each session. Lunch was provided between 12 PM and 1 PM for both AM and PM interviewees and interviewers.

Interviewers were requested to clear their schedules of clinical and research duties during interview days but were allowed the lunch break to handle emergencies as they arose.

Each interview panel used standardized questions to guide the interview but were allowed to ask follow-up questions as dictated by the flow of the interview. Interviews were graded on a scale of 0 to 5 with 5 being the best in the categories of academic record, professionalism, leadership, trainability, and fit for program for a total of 30 points.

Data Analysis

The raw interview scores were standardized to adjust for differences between interviewers and interview years by conversion to z-scores and percentile scores (0–100 scale). Normalization of data was confirmed quantitatively using the Shapiro–Wilk's test, and visually using density and Q–Q

plots. The standardized scores were sorted into categories to make the following comparisons: (1) different interview days (Day 1 vs. Day 2), (2) morning versus afternoon interviews, (3) different interview sessions (Day 1 AM/PM vs. Day 2 AM/PM), (4) before versus after breaks (morning break, lunch break, and afternoon break), (5) male versus female candidates, and (6) male versus female interviewers. Secondary analysis of data controlled for the presence of waitlisted candidates, studied the effects of candidate's fatigue, compared candidates who matched at Columbia versus all other candidates, and analyzed interviewer-specific characteristics (stage of career and experience). Student's *t*-test and analysis of variance were used for comparisons of continuous variables. Fisher's exact test was used for comparisons of categorical variables. Statistical analysis was conducted using R (version 3.6.3).

Results

A total of 387 candidates were interviewed from 2013 to 2019, with 183 male (47.3%) and 204 female (52.7%) candidates. Total 4,562 evaluations were completed by 40 interviewers: 19 male interviewers (47.5%) and 21 female interviewers (52.5%). The mean raw and standardized evaluation scores for each year are summarized in ►**Table 1**. On average, 27.6 candidates (standard deviation [SD] 5.1, range 20–36) per day and 13.8 candidates (SD 2.7, range 10–19) per session were interviewed.

Interview Day

There were no significant differences found between standardized evaluation scores received by candidates on interview Day 1 compared with Day 2 (51.51 vs. 50.45, $p = 0.22$), with similar scores observed regardless of interviewer's gender (►**Table 2**).

Time of Day

Morning interviews were defined as occurring before lunch, starting at 8:00 to 8:30 AM and ending before noon. Afternoon interviews were conducted after lunch, starting at 1:00 PM and ending before 5:00 PM. Percentile scores were found to be significantly higher in the morning compared with afternoon for all interviewers (52.75 vs. 49.28, $p < 0.001$), female interviewers (52.72 vs. 49.83, $p = 0.02$), and male interviewers (52.78 vs. 48.83, $p < 0.001$).

The data were further grouped to compare scores from early morning, late morning, early afternoon, and late afternoon cohorts. Early morning was defined as the first four interview slots of the day, late morning as the last four slots before lunch, early afternoon as the first four slots after lunch, and late afternoon as the last four slots of the day. Compared with candidates interviewed in the late afternoon, candidates in the early morning (54.47 vs. 46.74, $p < 0.001$), late morning (53.01 vs. 46.74, $p < 0.001$), and early afternoon (52.15 vs. 46.74, $p < 0.001$) received significantly higher mean percentile scores across all interviewers (►**Fig. 1A**), with the same patterns observed for both male and female interviewers (►**Table 2**). There were no significant differences between early morning, late morning, and early afternoon scores (►**Fig. 1A**) (►**Table 2**). The trend over the course of the interview day demonstrated a precipitous drop in interview scores for the late afternoon cohort rather than a gradual decline throughout the day.

Interview Session

Morning and afternoon sessions were divided into blocks occurring on different days, with comparisons made between interview Day 1 morning, Day 1 afternoon, Day 2 morning, and Day 2 afternoon sessions. Day 1 morning scores were higher than Day 1 afternoon (52.79 vs. 50.19, $p = 0.02$) and Day 2 afternoon scores (52.79 vs. 48.05, $p < 0.001$). Similarly, Day 2 morning scores were higher than Day 1 afternoon (52.71 vs. 50.19, $p = 0.03$) and Day 2 afternoon (52.71 vs. 48.05, $p < 0.001$) scores. No differences were observed between Day 1 morning and Day 2 morning scores (52.79 vs. 52.71, $p = 0.94$). These results suggest that morning scores from Day 1 and Day 2 were higher than afternoon scores from Day 1 and Day 2 (►**Fig. 1B**).

Before and After Breaks

Scores from the last four interviews before each break were compared with the scores from the first four interviews after

Table 1 Summary of candidate, interviewer, and evaluation characteristics for each interview year

	Interview year							
	2013	2014	2015	2016	2017	2018	2019	2013–2019
Number of candidates	63	61	46	50	60	53	54	387
Men	32 (50.8%)	26 (42.6%)	20 (43.5%)	28 (56.0%)	25 (41.7%)	23 (43.4%)	29 (53.7%)	183 (47.3%)
Women	31 (49.2%)	35 (57.4%)	26 (56.5%)	22 (44.0%)	35 (58.3%)	30 (56.6%)	25 (46.3%)	204 (52.7%)
Number of interviewer	13	14	11	11	11	16	15	40
Men	6 (46.2%)	7 (50.0%)	8 (72.7%)	7 (63.6%)	7 (63.6%)	9 (56.3%)	6 (40.0%)	19 (47.5%)
Women	7 (53.8%)	7 (50.0%)	3 (27.3%)	4 (36.4%)	4 (36.4%)	7 (43.7%)	9 (60.0%)	21 (52.5%)
Number of evaluations	784	739	449	534	649	732	675	4562
Average raw score	68.24	74.47	71.62	75.48	76.73	74.38	70.50	70.83
Average standardized percentile	53.48	54.01	52.16	54.56	54.76	51.63	51.73	51.05

Table 2 Mean standardized percentile scores for candidates organized by five key factors considered in our analysis and the gender of the interviewer

Factor		Gender of interviewers			*p*-Value
		Men	Women		
Interview day					
Thursday (Day 1)		51.63 (±1.47)	51.35 (±1.69)		0.81
Friday (Day 2)		49.76 (±1.76)	51.27 (±1.85)		0.25
		$p = 0.11$	$p = 0.95$		
Time of day					
Early AM		54.87 (±3.02)	53.96 (±3.47)		0.70
Late AM		52.33 (±2.87)	53.94 (±3.10)		0.46
	$p < 0.001$	$p = 0.23$	$p = 0.99$	**$p = 0.01$**	
Early PM		52.24 (±3.00)	52.03 (±3.26)		0.93
Late PM		45.99 (±2.99)	47.68 (±3.34)		0.46
		$p = 0.004$	$p = 0.067$		
Interview session					
Day 1 AM		53.00 (±2.07)	52.23 (±2.65)		0.65
Day 1 PM		50.22 (±2.10)	50.18 (±2.58)		0.98
		$p = 0.064$	$p = 0.28$		
Day 2 AM		52.49 (±2.41)	53.94 (±2.87)		0.44
Day 2 PM		46.93 (±2.56)	50.08 (±2.09)		0.061
		$p = 0.0019$	**$p = 0.033$**		
Interview breaks					
Before AM break		52.51 (±3.09)	50.55 (±3.51)		0.41
After AM break		51.78 (±2.96)	54.13 (±3.35)		0.30
		$p = 0.74$	$p = 0.15$		
Before lunch break		52.50 (±2.89)	53.65 (±3.14)		0.60
After lunch break		52.24 (±3.00)	52.03 (±3.26)		0.93
		$p = 0.90$	$p = 0.48$		
Before PM break		50.66 (±3.05)	49.96 (±3.45)		0.76
After PM break		47.39 (±3.07)	49.45 (±3.38)		0.38
		$p = 0.14$	$p = 0.84$		
Candidate's gender					
Female candidate		51.41 (±1.53)	51.73 (±1.75)		0.84
Male candidate		50.20 (±1.68)	50.85 (±1.78)		0.71
		$p = 0.30$	$p = 0.49$		

Notes: The 95th percentile confidence intervals are in parentheses. Comparisons between columns (male and female interviewers) have *p*-values in the far-right column, while comparisons between rows based on interview day factors have *p*-values listed in the row below. Significant *p*-values <0.05 are bolded for clarity.

the break. Break periods did not have a significant effect on interview scores. There was no difference in scores before versus after morning breaks (51.71 vs 52.83, $p = 0.5$), lunch breaks (53.01 vs. 52.15, $p = 0.6$), and afternoon breaks (50.35 vs, 48.30, $p = 0.2$).

Candidate's and Interviewer's Gender

There were no differences between the percentile scores given to female versus male candidates by all interviewers (51.55 vs. 50.49, $p = 0.2$), female interviewers (51.73 vs. 50.85, $p = 0.5$), and male interviewers (51.41 vs. 50.20, $p = 0.3$). No differences were found in scores given by female versus male interviewers in evaluating all candidates (51.31 vs. 50.84, $p = 0.6$).

Waitlisted Candidates

We performed a secondary analysis on the effects of interview day, time of day, and interview session for

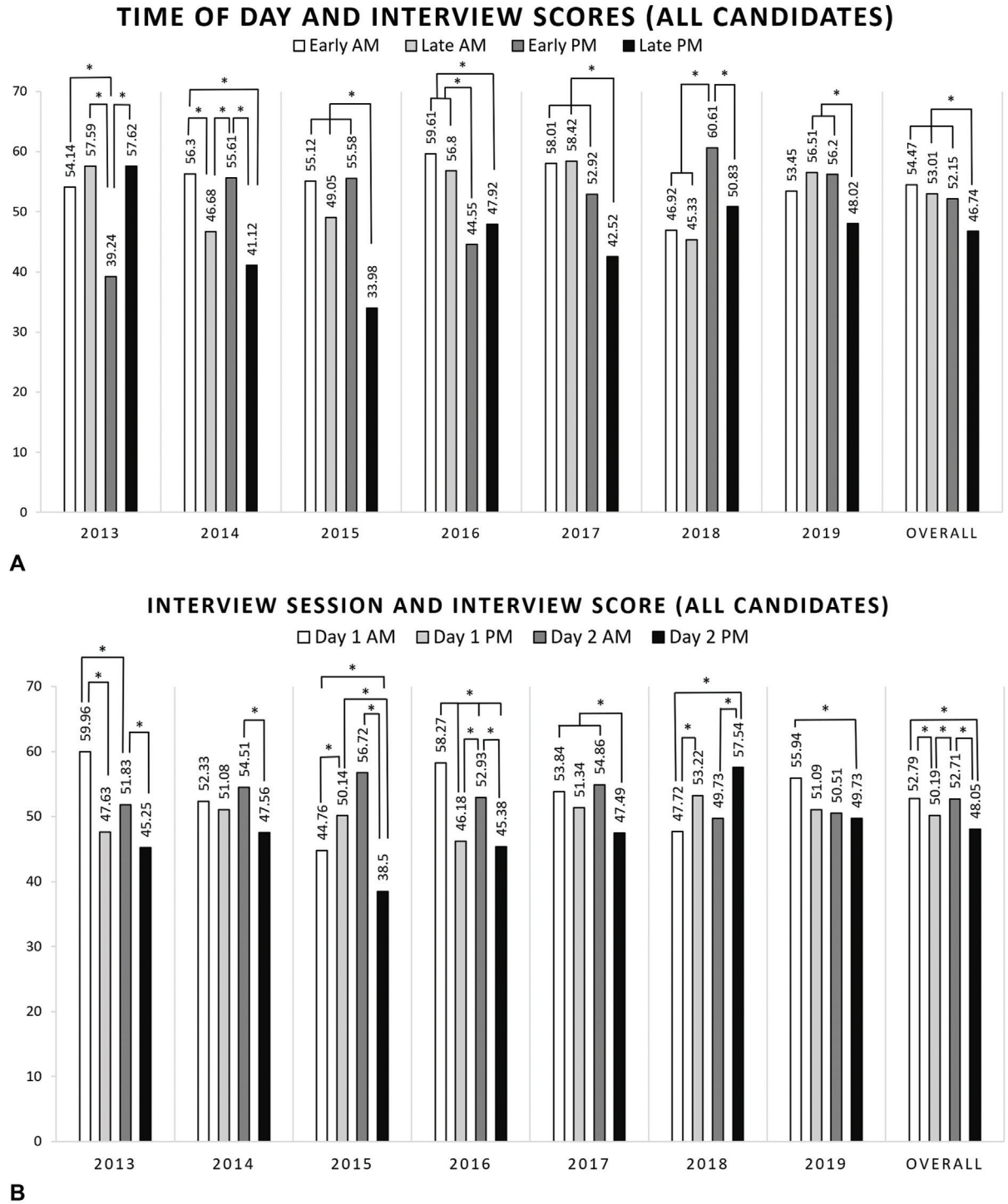

Fig. 1 Bar plot comparing the average standardized interview scores in **(A)** early AM (morning), late AM, early PM (afternoon), and late PM slots and **(B)** Day 1 AM, Day 1 PM, Day 2 AM, and Day 2 PM sessions for each individual year and overall (2013–2019). Significant differences with a *p*-value < 0.05 are indicated by asterisks.

nonwaitlisted candidates to control for the presence of waitlisted candidates as a potential confounder. In our study cohort, waitlisted candidates received lower average scores than nonwaitlisted candidates (44.29 vs. 52.29, $p < 0.001$). A total of 63 waitlisted candidates (17.6%) were interviewed between 2013 and 2019, with a range of 2 to 14 per year. These waitlisted candidates were offered unfilled interview spots, with the majority (54.0%) being placed in afternoon slots (►**Table 3**).

There were no differences noted in interview Day 1 versus Day 2 scores for nonwaitlisted candidates (51.97 vs. 52.46, $p = 0.6$). Significant differences remained between early morning (55.97 vs. 47.88, $p < 0.001$), late morning (53.44 vs. 47.88, $p = 0.001$), and early afternoon (53.31 vs. 47.88,

Table 3 Summary of waitlisted candidate characteristics including gender, interview session, and average standardized score compared with nonwaitlisted candidates

	Interview year							
	2013	2014	2015	2016	2017	2018	2019	2013–2019
Number of waitlisted candidates	14 of 63 (22.2%)	11 of 61 (18.0%)	12 of 46 (26.1%)	12 of 50 (24.0%)	9 of 60 (15.0%)	3 of 53 (5.7%)	2 of 54 (3.7%)	63 of 357 (17.6%)
Men	6 (42.9%)	4 (36.4%)	7 (58.3%)	5 (41.7%)	4 (44.4%)	1 (33.3%)	1 (50.0%)	28 (44.4%)
Women	8 (57.1%)	7 (63.6%)	5 (41.7%)	7 (58.3%)	5 (55.6%)	2 (66.7%)	1 (50.0%)	35 (55.6%)
Interview session								
Day 1 AM	5 (35.7%)	2 (18.2%)	3 (25.0%)	0	3 (33.3%)	0	0	13 (20.6%)
Day 1 PM	5 (35.7%)	1 (9.1%)	1 (8.3%)	2 (16.7%)	1 (11.1%)	0	0	10 (15.9%)
Day 2 AM	1 (7.2%)	3 (27.3%)	3 (25.0%)	4 (33.3%)	2 (22.2%)	3 (100.0%)	0	16 (25.4%)
Day 2 PM	3 (21.4%)	5 (45.5%)	5 (41.7%)	6 (50.0%)	3 (33.3%)	0	2 (100.0%)	24 (38.1%)
Average standardized percentile—waitlist	43.02	44.51	42.97	47.26	45.43	32.01	57.68	44.29
Average standardized percentile—nonwaitlist	53.27	52.90	49.31	52.05	52.83	52.79	51.50	52.29

$p = 0.002$) evaluation scores when compared with late afternoon, which demonstrated that the drop in late afternoon scores persisted despite controlling for the presence of waitlisted candidates (►**Fig. 2A**). Day 1 morning scores remained higher than Day 2 afternoon scores (53.28 vs. 50.66, $p = 0.03$), and Day 2 afternoon scores (53.28 vs. 50.49, $p = 0.04$), while Day 2 morning scores also remained higher than Day 1 afternoon (54.66 vs. 50.66, $p = 0.002$) and Day 2 afternoon scores (54.66 vs. 50.49, $p = 0.005$) (►**Fig. 2B**). There was blunting of the decline in afternoon scores when controlling for the presence of waitlisted candidates (►**Fig. 3**) (►**Table 4**), but the difference between morning and afternoon scores remained statistically significant.

Number of Candidates Interviewed

The number of candidates interviewed per session and per day ranged from 10 to 19 and 20 to 36, respectively. There was poor correlation between number of candidates per session and mean interview scores for the session (Pearson's $r = 0.01$) and between the number of candidates interviewed per day and afternoon interview scores (Pearson's $r = 0.09$).

Candidate's Fatigue

Secondary ad hoc outcomes of candidate's fatigue were measured to evaluate an alternative explanation for the lower afternoon scores observed in the study cohort. There was no significant correlation between interview session position, defined as the number of students who were interviewed before the candidate, and the interview score (Pearson's $r = -0.07$). In addition, there were no differences noted between the first two interviews and the last two interviews for all candidates (51.44 vs. 50.84, $p = 0.54$), morning candidates (53.08 vs. 53.22, $p = 0.92$), and afternoon candidates (49.76 vs. 48.25, $p = 0.28$), which suggested that progressively poor candidate performance in later interviews was an unlikely independent explanation for the decline in afternoon interview scores.

Candidates Matching at Columbia

Comparisons were also made regarding interview time of day between candidates who matched at the Columbia residency program compared with all other candidates. Out of 23 candidates who matched at Columbia, 9 (39.1%) were interviewed in the morning and 14 (60.9%) in the afternoon, and for the other 318 candidates, 166 (52.2%) were interviewed in the morning and 152 (47.8%) in the afternoon. There were no differences in the proportion of Columbia matched versus all other candidates in morning and afternoon sessions ($p = 0.23$). This suggests that although time-of-day effects were seen on interview scores, they may not have significantly impacted the candidates who matched with the program. While it is expected that candidates receiving the highest interview scores would be less affected by interview day factors, our study design and data collection are limited in the ability to discern the possibility for the evaluation of other equally top candidates to be affected by time-of-day effects.

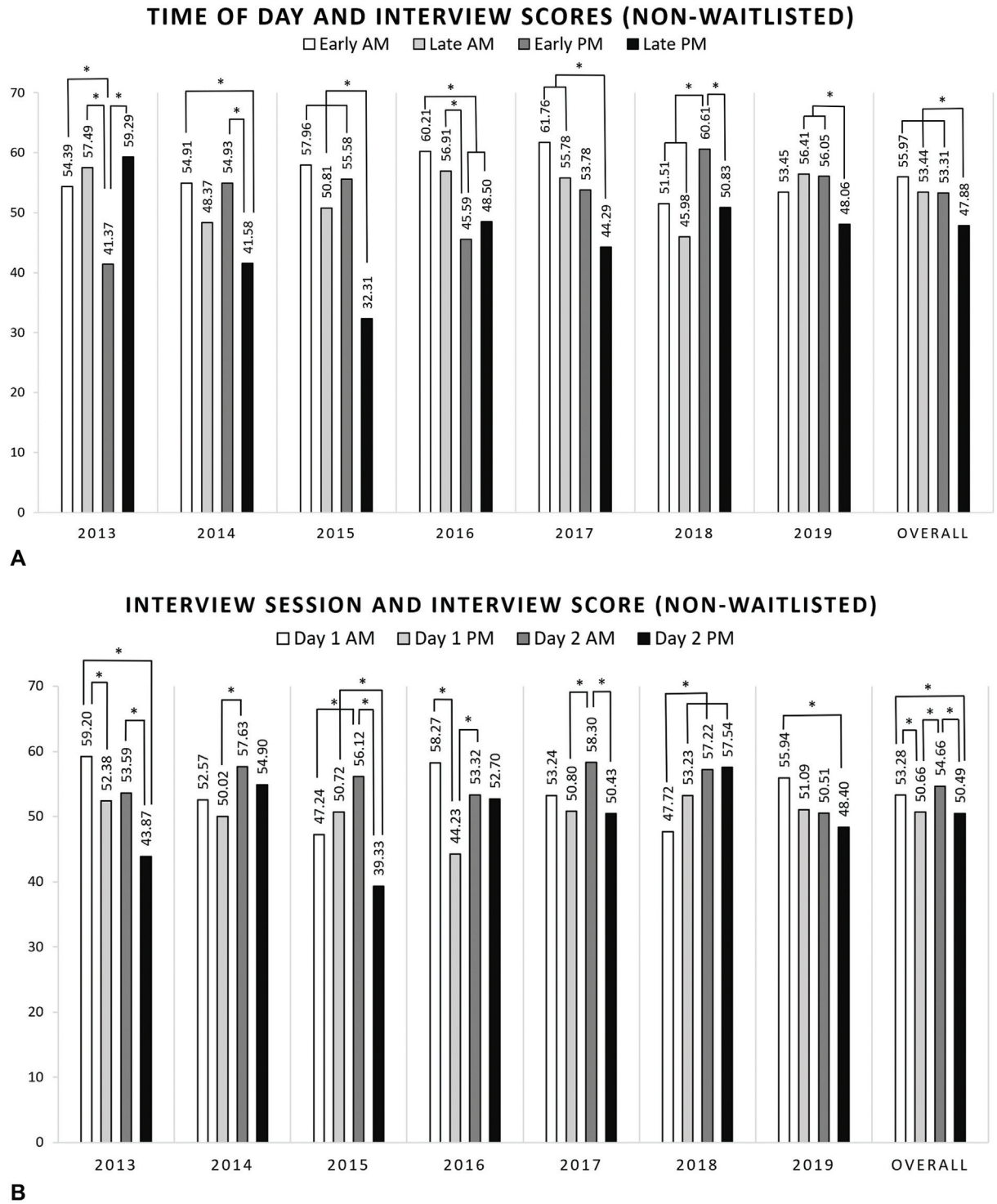

Fig. 2 To control for the presence of waitlisted candidates, we performed secondary analysis of time-of-day effects on interview scores with nonwaitlisted candidates. The bar plots are comparing the average standardized interview scores in **(A)** early AM (morning), late AM, early PM (afternoon), and late PM slots and **(B)** Day 1 AM, Day 1 PM, Day 2 AM, and Day 2 PM sessions for each individual year and overall (2013–2019). Significant differences with a p-value < 0.05 are indicated by asterisks.

Interviewer's Characteristics

Finally, we explored whether interviewer-specific characteristics, including stage of career and experience of interviewer, were independently associated with time-of-day and interview session effects on evaluation scores. Only scores obtained from nonwaitlisted candidates were considered to distinguish interviewer's factors from waitlist-related differences. Interviewers were categorized as chief residents, junior faculty, mid-career faculty, and senior faculty to represent stages of career. In addition, interviewers were sorted into two categories based on resident selection experience: 4 or more years of experience conducting residency

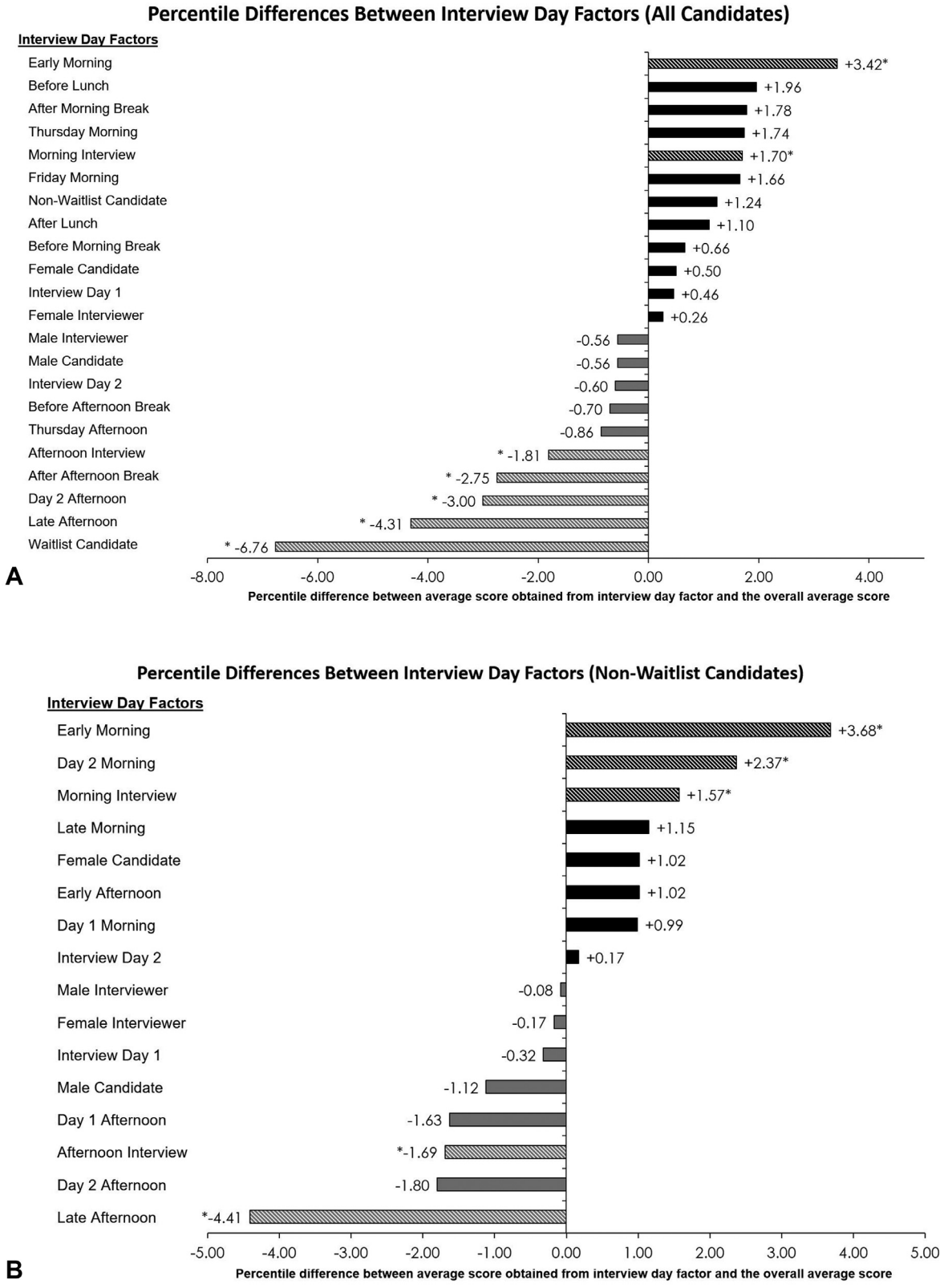

Fig. 3 Diverging bar plot with numerical values representing percentile differences between average standardized interview scores obtained from the multiple listed interview day factors and the overall average standardized interview score for **(A)** all candidates and **(B)** nonwaitlisted candidates. Positive percentile differences are interpreted as average scores that are higher than the overall average standardized interview score, with statistically significant differences ($p < 0.05$) indicated by a black-and-white diagonal pattern and an asterisk, while nonsignificant differences are in black. Negative percentile differences are interpreted as average scores that are lower than the overall average standardized interview score, with statistically significant differences ($p < 0.05$) indicated by a gray-and-white diagonal pattern and an asterisk, while nonsignificant differences are in gray. Compared with **(A)**, when we adjusted for the presence of waitlisted candidates in **(B)**, the difference between early morning and late afternoon scores persisted.

Table 4 Mean standardized percentile ratings for candidates organized by five key factors considered in our analysis and the type of candidate

Factor		Type of candidate			*p*-Value
		Nonwaitlisted	All candidates		
Interview day					
Thursday (Day 1)		51.97 (±1.47)	51.51 (±1.11)		0.57
Friday (Day 2)		52.46 (±1.76)	50.45 (±1.27)		**0.04**
		$p = 0.60$	$p = 0.22$		
Time of day					
Early AM		55.97 (±2.42)	54.47 (±2.27)		0.38
Late AM		53.44 (±2.23)	53.01 (±2.12)		0.79
	$p < 0.001$	$p = 0.13$	$p = 0.36$	$p < 0.001$	
Early PM		53.31 (±2.34)	52.15 (±2.20)		0.48
Late PM		47.88 (±2.53)	46.74 (±2.22)		0.51
		$p = 0.002$	$p = 0.0007$		
Interview session					
Day 1 AM		53.28 (±1.65)	52.79 (±1.57)		0.67
Day 1 PM		50.66 (±1.65)	50.19 (±1.57)		0.68
		$p = 0.028$	$p = 0.022$		
Day 2 AM		54.66 (±1.89)	52.71 (±1.74)		0.14
Day 2 PM		50.49 (±2.18)	48.05 (±1.86)		0.094
		$p = 0.0046$	$p = 0.0003$		
Candidate's gender					
Female candidate		53.31 (±1.25)	51.55 (±1.15)		**0.042**
Male candidate		51.17 (±1.31)	50.49 (±1.22)		0.46
		$p = 0.02$	$p = 0.21$		
Interviewer's gender					
Female interviewer		52.12 (±1.34)	51.31 (±1.25)		0.39
Male interviewer		52.21 (±1.22)	50.84 (±1.13)		0.11
		$p = 0.92$	$p = 0.58$		

Notes: The 95th percentile confidence intervals are in parentheses. Comparisons between columns based on type of candidate have *p*-values in the far-right column, while comparisons between rows based on interview day factors have *p*-values in the row below. Significant *p*-values < 0.05 are bolded for clarity.

interviews versus 3 or fewer years of experience. The distinction between 3 and 4 years of experience was determined to minimize differences in the sample size.

In the evaluation of time-of-day effects, we found that junior, mid-career, and senior faculties all gave higher interview scores in either the early or late morning sessions compared with late afternoon (►**Fig. 4A**). There was a similar degree of higher scores observed for chief residents in late morning (55.79 vs. 48.45, $p = 0.07$) compared with late afternoon, but potentially missing significance due to a lower number of total evaluations completed by chief residents ($n = 365$) compared with other faculty ($n = 541$–830). Both more experienced (4 or more years of selection experience) and less experienced (3 or fewer years) faculty interviewers gave higher scores in the early morning, late morning, and early afternoon compared with late afternoon (►**Fig. 4A**). Our results suggested that interviewers of any stage of career or experience level were susceptible to time-of-day effects on interview scores, with lowest mean evaluation scores found in the late afternoon cohort.

In the assessment of interview sessions, we found mid-career faculty, senior faculty, and interviewers with 4 or more years of experience associated with lower scores in Day 2 afternoon in comparison to Day 1 morning. In the senior faculty cohort, Day 2 afternoon scores were found to be lower than all other times of the day (44.74 vs. 52.03, 49.74, 52.57, $p < 0.001$). Interviewers with 4 or more years of experience also gave lower scores in Day 1 afternoon compared with Day 1 morning (49.50 vs. 52.91, $p = 0.03$). There were no differences between interview session scores for chief residents, junior faculty, and interviewers with 3 or fewer years of experience (►**Fig. 4B**).

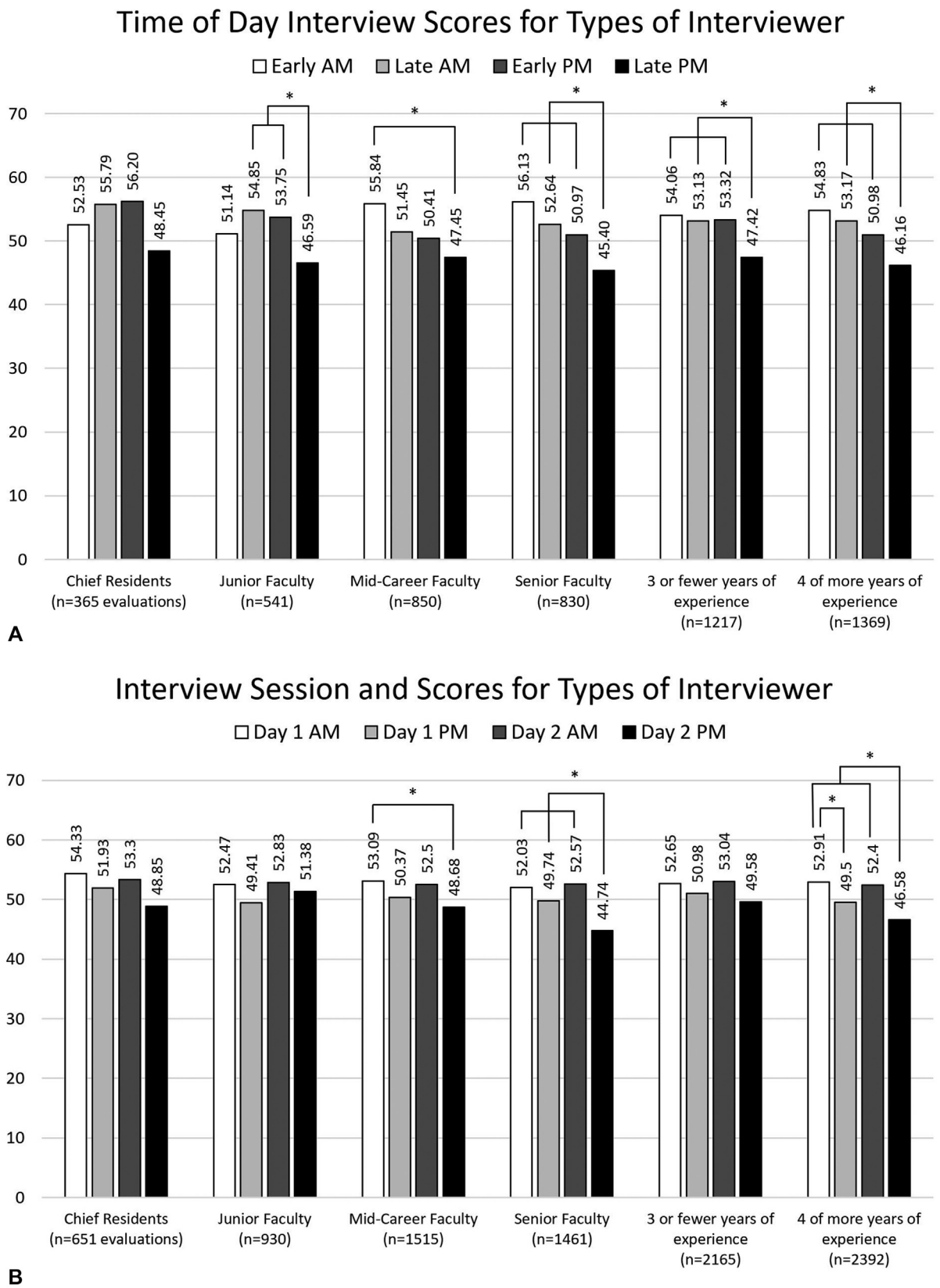

Fig. 4 Bar plot comparing the average standardized interview scores in **(A)** early AM (morning), late AM, early PM (afternoon), and late PM slots and **(B)** Day 1 AM, Day 1 PM, Day 2 AM, and Day 2 PM sessions for interviewer groups separated by stage of career (chief residents, junior faculty, mid-career faculty, and senior faculty) and years of residency recruitment experience. The number of evaluations submitted by each interviewer group is listed under the category. Significant differences with a p-value < 0.05 are indicated by asterisks.

Discussion

Interviews are a critical component to the successful matching of applicants into residency programs. Significant efforts to reduce interrater variability have been made to standardize the interview and create a fair, transparent process.[28] For example, residency programs have created checklists of desired traits, utilized standardized questions, adopted grading scales with a defined rubric, used multiple interviewers, and blinded the interviewer to application data.[29] Although the process has become more systematic, a busy all-day interview schedule may generate fluctuations in cognitive and emotional energy in faculty interviewers and/or candidates who can influence the scores given throughout the day. There is a paucity of prior literature studying time-of-day effects on interview performance, with only a single study at an electric utility company that showed no appreciable effect on interview ratings.[30]

In our study of a 7-year residency interview cohort for the ophthalmology residency program at a single academic medical center, we found significant time-of-day effects on interview scores, where candidates received higher scores in the morning compared with the afternoon, with the lowest scores identified in the late afternoon cohort. We propose that the lower scores observed in the late afternoon may be related to increased cognitive load and decision-making fatigue as faculty members balance clinical and research responsibilities while managing an all-day interview process. Fatigue at the end of the day has been associated with impairments in accurate recall,[31] decision-making,[32] social judgments, short- and long-term memories, and may predispose interviewers to develop cognitive shortcuts[33] that contribute to an inaccurate and poorer representation of the candidate's performance. In addition, prior studies have suggested an increased sensitivity to emotionally negative stimuli later in the day, largely contributed by fatigue exacerbating feelings of anxiety and decreased mood[34] which can result in more critical evaluations. Candidates at the end of the day may also become more susceptible to narrow bracketing practices, a phenomenon demonstrated over 10 years of interview data from business programs.[35] In narrow bracketing, interviewers attempt to minimize deviation from the expected distribution of scores, which means if the interviewers were giving higher scores in the morning, they often became more critical of applicants in the afternoon. Even if this practice yields more accurate scores at the end of the day, it may also disadvantage those candidates. These are all important factors to consider as time-of-day effects on decision-making, emotional judgment, recall, and narrow bracketing can inaccurately and unfairly represent candidate's performance.

The impact of fatigue, the circadian rhythm, and the end-of-day phenomenon on cognitive tasks and job performance have been suggested in a broad range of literature, including colonoscopy detection rates,[36] risk for neonatal death,[37] incidence of anesthesia adverse events,[38] patient morbidity and mortality after surgery,[39] errors in navigation,[40] driving performance,[41] and judicial decision-making.[42] In these studies, the two explanations most discussed for decreased task performance were mental fatigue and ego depletion. Ego depletion is defined as a reduction in a person's willingness to complete complex tasks or make decisions after expending self-control in a previous task. The classic example of ego depletion is described in the study of judicial rulings by Danziger et al,[42] where the percentage of favorable rulings was found to decline from 65% to nearly zero and then returned to 65% following food breaks. The argument was that judges experienced a gradual depletion of mental resources that led to unfavorable decisions and decision-making capacity returned to baseline following breaks. However, the problem with the ego-depletion argument is the presence of external factors that may provide alternative explanations. For example, subsequent analyses revealed that the nonrandom order of judicial cases resulted in a greater number of unrepresented prisoners scheduled at the end of sessions[43] and allocated unfavorable cases that would take less time before breaks.[44] When controlling for these factors, the ego-depletion effect was reduced significantly. In addition, meta-analyses of ego-depletion studies showed significant bias from small-study effects and publication bias,[45] medium to high heterogeneity among studies,[46] and no significant effects observed during replication experiments[47] or when controlling for variabilities in effect size.[48]

To avoid these pitfalls, we considered multiple external factors that could modify the time-of-day effects that we observed on interview scores. We identified that waitlisted candidates received lower interview scores compared with nonwaitlisted candidates and would potentially be relegated to less desirable interview times and days, but the drop in late afternoon scores persisted when waitlisted candidates were removed from the analysis. We also considered candidate's fatigue as an external factor but found no significant correlation between candidate's position and interview score along with no differences in score between the first and final interview sessions for candidates. This analysis, however, only considers candidates' performances during 1-to-2-hours window in which they were interviewed and may not be applicable to the entire day. Unfortunately, we did not have data spanning the whole day for individual candidates as we had for interviewers. Finally, we assessed whether time-of-day effects could be explained by specific interviewer's characteristics. While lower late afternoon interview scores were observed in nearly all interview groups regardless of stage of career or years of interview experience, they were more significant in mid-career and senior faculty as well as those with more experience.

Previous studies have suggested that female candidates may be subject to more negative interview experiences compared with male candidates with gender-specific inappropriate behavior or illegal questions,[49,50] although a separate study suggested that female candidates received higher interview scores.[16] We were reassured to find no appreciable candidate's and interviewer's gender biases in our process but must continue to remain vigilant. Across all interview years, the gender proportion of interviewers and interviewees were closely matched, which we hypothesize may

contribute toward reducing gender-related biases. Although interview scores were found to be lower for candidates in the afternoon, there were no differences in time of day for candidates who ultimately matched into our program. We have implemented changes to our interview schedules which should mitigate the time-of-day effects this study uncovered. Moving forward, we will continue to monitor the effects of these interventions in the Plan-Do-Study-Act cycle of continuous quality improvement.

Due to the coronavirus disease 2019 pandemic, ophthalmology residency interviews have been conducted virtually beginning in 2020. Increased fatigue has been associated with virtual meetings in the workplace.[51] Researchers cite several factors for this added fatigue including excessive amounts of close-up eye gaze, increased self-evaluation from staring at the video of oneself, constraints on physical mobility, and increased cognitive load.[52] It is unclear how the virtual format will affect end-of-day fatigue in residency interviews.

Limitations of our study include data being collected from the residency interview cohort at a single academic medical center for a single surgical specialty. Our findings may not be applicable to other institutions and specialties. The structure of the interview day, grading practices, and type of interviews are just a few of many factors which may be different between individual residency programs. For example, the number of candidates interviewed per day ranged between 20 and 36, which may be high compared with other institutions. Although we saw no correlation between number of candidates interviewed and scores, it is impossible to determine if this would be true for sessions containing significantly fewer or significantly greater numbers of candidates based on our data. In addition, our interview day structure did not allow us to determine if candidate's fatigue may also play a role, although we found that it did not vary during the span of their five to six interviews. Finally, the effects of other factors such as race, ethnicity, sexual orientation, and candidate's age were not studied due to limitations in the collected retrospective data.

Conclusion

Residency interview scores may be influenced by the time of day, a factor that is relevant to the interview process for most medical and surgical specialties. While this effect appears more prominently among more experienced and senior faculties, all graders were susceptible. We recommend future studies to replicate our findings for other residency programs in ophthalmology and in other specialties to evaluate whether adjusting the structure of the interview day is warranted to improve the fairness of the interview process for candidates. Future studies should also focus on the effects of candidate's race, ethnicity, sexual orientation, and age.

Ethical Approval

The project was approved by the Columbia University Irving Medical Center Institutional Review Board (IRB #AAAT2246) on September 8, 2020, and was compliant with protection of individually identifiable information.

Disclaimers

None.

Meeting Presentation

Association of University Professors of Ophthalmology (AUPO) Annual Meeting, February 4–6, 2021, virtual.

Funding/Support

This research was supported, in part, by the UCSF Vision Core shared resource of the NIH/NEI P30 EY002162 and unrestricted departmental grants from Research to Prevent Blindness to the Department of Ophthalmology at Columbia University and UCSF.

Conflict of Interest

None declared.

Acknowledgments

The authors wish to thank the residents, faculty, and staff of the Columbia University Medical Center's Department of Ophthalmology that helped facilitate residency interviews and the evaluation of residency candidates.

References

1 Taylor CA, Weinstein L, Mayhew HE. The process of resident selection: a view from the residency director's desk. Obstet Gynecol 1995;85(02):299–303
2 Fine PL, Hayward RA. Do the criteria of resident selection committees predict residents' performances? Acad Med 1995;70(09): 834–838
3 Green M, Jones P, Thomas JX Jr. Selection criteria for residency: results of a national program directors survey. Acad Med 2009;84 (03):362–367
4 Westerman ME, Boe C, Bole R, et al. Evaluation of medical school grading variability in the United States: are all honors the same? Acad Med 2019;94(12):1939–1945
5 Lee AG, Golnik KC, Oetting TA, et al. Re-engineering the resident applicant selection process in ophthalmology: a literature review and recommendations for improvement. Surv Ophthalmol 2008; 53(02):164–176
6 Davis JL, Platt LD, Sandhu M, Shapiro F. Evaluating factors in the selection of residents. Acad Med 1995;70(03):176–177
7 Sabin J, Nosek BA, Greenwald A, Rivara FP. Physicians' implicit and explicit attitudes about race by MD race, ethnicity, and gender. J Health Care Poor Underserved 2009;20(03):896–913
8 Capers Q IV, Clinchot D, McDougle L, Greenwald AG. Implicit racial bias in medical school admissions. Acad Med 2017;92(03): 365–369
9 Osseo-Asare A, Balasuriya L, Huot SJ, et al. Minority resident physicians' views on the role of race/ethnicity in their training experiences in the workplace. JAMA Netw Open 2018;1(05): e182723
10 Smith CJ, Rodenhauser P, Markert RJ. Gender bias of Ohio physicians in the evaluation of the personal statements of residency applicants. Acad Med 1991;66(08):479–481
11 Filippou P, Mahajan S, Deal A, et al. The presence of gender bias in letters of recommendations written for urology residency applicants. Urology 2019;134:56–61

12 Lin F, Oh SK, Gordon LK, Pineles SL, Rosenberg JB, Tsui I. Gender-based differences in letters of recommendation written for ophthalmology residency applicants. BMC Med Educ 2019;19(01):476
13 Turrentine FE, Dreisbach CN, St Ivany AR, Hanks JB, Schroen AT. Influence of gender on surgical residency applicants' recommendation letters. J Am Coll Surg 2019;228(04):356–365.e3
14 Grimm LJ, Redmond RA, Campbell JC, Rosette AS. Gender and racial bias in radiology residency letters of recommendation. J Am Coll Radiol 2020;17(1 Pt A):64–71
15 Rand VE, Hudes ES, Browner WS, Wachter RM, Avins AL. Effect of evaluator and resident gender on the American Board of Internal Medicine evaluation scores. J Gen Intern Med 1998;13(10): 670–674
16 Oyler J, Thompson K, Arora VM, Krishnan JA, Woodruff J. Faculty characteristics affect interview scores during residency recruitment. Am J Med 2015;128(05):545–550
17 Loeppky C, Babenko O, Ross S. Examining gender bias in the feedback shared with family medicine residents. Educ Prim Care 2017;28(06):319–324
18 Nallasamy S, Uhler T, Nallasamy N, Tapino PJ, Volpe NJ. Ophthalmology resident selection: current trends in selection criteria and improving the process. Ophthalmology 2010;117(05): 1041–1047
19 Wagoner NE, Suriano JR, Stoner JA. Factors used by program directors to select residents. J Med Educ 1986;61(01):10–21
20 Brothers TE, Wetherholt S. Importance of the faculty interview during the resident application process. J Surg Educ 2007;64(06): 378–385
21 Brustman LE, Williams FL, Carroll K, Lurie H, Ganz E, Langer O. The effect of blinded versus nonblinded interviews in the resident selection process. J Grad Med Educ 2010;2(03):349–353
22 Hauge LS, Stroessner SJ, Chowdhry S, Wool NLAssociation for Surgical Education. Evaluating resident candidates: does closed file review impact faculty ratings? Am J Surg 2007;193(06): 761–765
23 Maxfield CM, Thorpe MP, Desser TS, et al. Bias in radiology resident selection: do we discriminate against the obese and unattractive? Acad Med 2019;94(11):1774–1780
24 Corcimaru A, Morrell MC, Morrell DS. Do looks matter? The role of the Electronic Residency Application Service photograph in dermatology residency selection. Dermatol Online J 2018;24(04): 1–4
25 Heidemann DL, Thompson E, Drake SM. Does timing of internal medicine residency interview affect likelihood of matching? South Med J 2016;109(08):466–470
26 Martin-Lee L, Park H, Overton DT. Does interview date affect match list position in the emergency medicine national residency matching program match? Acad Emerg Med 2000;7(09): 1022–1026
27 Avasarala S, Thompson E, Whitehouse S, Drake S. Assessing correlation of residency applicants' interview dates with likelihood of matching. South Med J 2018;111(02):83–86
28 Neitzschman HR, Neitzschman LH, Dowling A. Key component of resident selection: the semistructured conversation. Acad Radiol 2002;9(12):1423–1429
29 Stephenson-Famy A, Houmard BS, Oberoi S, Manyak A, Chiang S, Kim S. Use of the interview in resident candidate selection: a review of the literature. J Grad Med Educ 2015;7(04):539–548
30 Willihnganz MA, Meyers LS. Effects of time of day on interview performance. Public Pers Manage 1993;22(04):545–550
31 Petros TV, Beckwith BE, Anderson M. Individual differences in the effects of time of day and passage difficulty on prose memory in adults. Br J Psychol 1990;81(01):63–72
32 Monk TH, Leng VC. Time of day effects in simple repetitive tasks: Some possible mechanisms. Acta Psychol (Amst) 1982;51(03): 207–221
33 Lewandowska K, Wachowicz B, Marek T, Oginska H, Fafrowicz M. Would you say "yes" in the evening? Time-of-day effect on response bias in four types of working memory recognition tasks. Chronobiol Int 2018;35(01):80–89
34 Gobin CM, Banks JB, Fins AI, Tartar JL. Poor sleep quality is associated with a negative cognitive bias and decreased sustained attention. J Sleep Res 2015;24(05):535–542
35 Simonsohn U, Gino F. Daily horizons: evidence of narrow bracketing in judgment from 10 years of M.B.A. admissions interviews. Psychol Sci 2013;24(02):219–224
36 Lee A, Iskander JM, Gupta N, et al. Queue position in the endoscopic schedule impacts effectiveness of colonoscopy. Am J Gastroenterol 2011;106(08):1457–1465
37 Pasupathy D, Wood AM, Pell JP, Fleming M, Smith GC. Time of birth and risk of neonatal death at term: retrospective cohort study. BMJ 2010;341:c3498
38 Wright MC, Phillips-Bute B, Mark JB, et al. Time of day effects on the incidence of anesthetic adverse events. Qual Saf Health Care 2006;15(04):258–263
39 Kelz RR, Tran TT, Hosokawa P, et al. Time-of-day effects on surgical outcomes in the private sector: a retrospective cohort study. J Am Coll Surg 2009;209(04):434–445.e2
40 Zhang X, Qu X, Xue H, Tao D, Li T. Effects of time of day and taxi route complexity on navigation errors: an experimental study. Accid Anal Prev 2019;125:14–19
41 Lenné MG, Triggs TJ, Redman JR. Time of day variations in driving performance. Accid Anal Prev 1997;29(04):431–437
42 Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. Proc Natl Acad Sci U S A 2011;108(17):6889–6892
43 Weinshall-Margel K, Shapard J. Overlooked factors in the analysis of parole decisions. Proc Natl Acad Sci U S A 2011;108(42):E833, author reply E834
44 Glockner A. The irrational hungry judge effect revisited: simulations reveal that the magnitude of the effect is overestimated. Judgm Decis Mak 2016;11(06):601–610
45 Carter EC, Kofler LM, Forster DE, McCullough ME. A series of meta-analytic tests of the depletion effect: self-control does not seem to rely on a limited resource. J Exp Psychol Gen 2015;144(04): 796–815
46 Dang J. An updated meta-analysis of the ego depletion effect. Psychol Res 2018;82(04):645–651
47 Hagger MS, Chatzisarantis NLD, Alberts H, et al. A multilab preregistered replication of the ego-depletion effect. Perspect Psychol Sci 2016;11(04):546–573
48 Friese M, Loschelder DD, Gieseler K, Frankenbach J, Inzlicht M. Is ego depletion real? An analysis of arguments. Pers Soc Psychol Rev 2019;23(02):107–131
49 Hern HG Jr, Alter HJ, Wills CP, Snoey ER, Simon BC. How prevalent are potentially illegal questions during residency interviews? Acad Med 2013;88(08):1116–1121
50 Lee JS, Ji YD, Kushner H, Kaban LB, Peacock ZS. Residency interview experiences in oral and maxillofacial surgery differ by gender and affect residency ranking. J Oral Maxillofac Surg 2019;77(11):2179–2195
51 Bennett AA, Campion ED, Keeler KR, Keener SK. Videoconference fatigue? Exploring changes in fatigue after videoconference meetings during COVID-19. J Appl Psychol 2021;106(03):330–344
52 McNamara , D. S., Bailenson , J. N. Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue. Technology, Mind, and Behavior 2021;2(01):. Doi: 10.1037/tmb0000030