Running Head: Daily Horizons

Daily Horizons: Evidence of Narrow Bracketing in Judgment from 10 years of MBA-admission Interviews

Uri Simonsohn
University of Pennsylvania
uws@wharton.upenn.edu

Francesca Gino
Harvard University
fgino@hbs.edu

Forthcoming at *Psychological Science*

Abstract

Many professionals, from auditors and lawyers to clinical psychologists and journal editors, divide a continuous flow of judgments into subsets. College admissions interviewers, for instance, evaluate just a handful of applicants per day. We conjectured that in such situations individuals engage in "narrow bracketing" –that is, assessing each subset in isolation and, for that subset, avoiding much deviation from the expected overall distribution of judgments. For instance, an interviewer who has already highly recommended three applicants on a given day may be reluctant to do so for a fourth applicant. Data from over 9,000 MBA interviews supported this prediction. Auxiliary analyses suggest that contrast effects and non-random scheduling of interviews are unlikely alternative explanations for the phenomenon.

Many professionals, from auditors and venture capitalists to clinical psychologists and journal editors, have jobs that involve a continuous flow of judgments that they execute, over time, in small subsets. University admissions officers, for example, interview hundreds of applicants per year in subsets of a handful each day.

These arbitrarily created subsets should have no influence on experts' judgments. While the merit of an MBA applicant may partially depend on the pool of applicants *that year*, it should not depend on the few others randomly interviewed *that day*. However, decision makers often engage in "narrow bracketing" –that is, they fail to integrate the consequences of many similar decisions into their judgments (for a review see Read, Loewenstein, & Rabin, 1999).

In this paper, we examine the phenomenon of narrow bracketing in judgment. In line with research showing that people focus too much on the particular case at hand and neglect background information (Brenner, Griffin, & Koehler, 2005; Griffin & Tversky, 1992; Massey & Wu, 2005), we propose that when people conduct a subset of judgments, they do not sufficiently consider the other subsets they have already made or will make in the future. Considering that people exaggerate the extent to which small samples resemble large samples (Tversky & Kahneman, 1971), we conjecture that people avoid making *subsets* of judgments that deviate much from what they expect the *overall* set of judgments to be like. For instance, an interviewer who expects to evaluate positively about 50% of applicants in a pool may be reluctant to evaluate positively many more or fewer than 50% of applicants on any given day. An applicant who happens to interview on a day when several others have already received a positive evaluation would, therefore, be at a disadvantage.

We tested this prediction by analyzing ten years of data of MBA applications to an American business school (to which neither author is affiliated), assessing whether applicants

interviewed following an above (below) average set of applicants that day, received lower (higher) scores.

We studied narrow bracketing in the contexts of experts working in their everyday environment rather than in the laboratory in order to avoid the possibility that judgments would be negatively autocorrelated due to knowledge of the distribution of underlying quality (e.g., "I rated the previous three applicants very highly, so the pool must be strong, and I will start evaluating more harshly") or scale use (e.g., "I am giving too many 4s, so I should give more 3s and reserve 4s for extraordinary applicants"). Because experts who interview large numbers of applicants year after year should not revise their beliefs or scale usage upon seeing a handful of weak or strong applicants on a single day, they should be an ideal testing ground for narrow bracketing.

Empirical Analyses

Data

The dataset consists of 14,065 interviews of MBA applicants between 2000 and 2009.[1] After erroneous or incomplete entries and interviews conducted by alumni were eliminated, the sample contained information on 9,323 interviews conducted by 31 interviewers. Following the disclosure guidelines by Simmons, Nelson, and Simonsohn (2011), supplemental materials contain the full list of variables. Given the nested nature of the data, all analyses cluster standard errors at the interviewer level (i.e., taking into account that we have repeated measures for interviewers).

Interviewers rate applicants on five subscores (in 1-5 scales): (i) communication skills, (ii) being driven, (iii) ability to work in teams, (iv) being accomplished, and (v) interest in the

[1] Applications for the academic year 2004-2005 are missing from our records.

school. They also provide an overall 1-5 evaluation on both the written application and the interview. We refer to the latter as "scores" and the former as "subscores."

Interviewers conduct an average of 4.5 interviews per day (SD=1.9), conditional on conducting at least one, and give an average score of 2.8 (SD=0.9). The dataset includes information both about the applicants (e.g., their GMAT scores) and the interview (e.g., date and time).

Main results

We estimated regressions with applicants' score as the dependent variable and the average score given to previous applicants by the same interviewer earlier that day as the key predictor. We control for applicant and interview characteristics and for interviewer fixed effects.

Column 1 in Table 1 reports the baseline specification, controlling only for interviewer effects (with 30 binary variables that allow estimation of a separate main effect for each of the 31 interviewers) and for month and year of the interview (allowing a main effect for each month of each year in the sample). The point estimate for the impact of the average score of previous interviews is, as predicted, negative and significant (B=-.116, p=.005).

Column 2 adds controls for applicant characteristics, and Column 3 adds controls for interview characteristics. The point estimate of interest is still negative and significant in both models. In Column 4, we add the (1-5) score given to an applicant's *written* application. This score should control for many other unobservable differences across applicants. The point estimate of interest remains negative and significant (B=-.088, *p*=.018).[2] The stability of the key

[2] A referee noted that regressions (with small samples) where a predictor is the average of multiple lags of the dependent variable can lead to spurious negative correlations. To assess if this was a problem in our data, we created 100 mock datasets randomly resorting the interviews by each interviewer across days and estimated the regressions reported in Table 1 (re-computing mock daily averages). The point estimate of interest was on average B=-.002,

point estimates when adding controls into the regression gives us confidence that the main finding is not the result of omitted variables.[3]

Figure 1 depicts the residuals from a regression that controls for all observables in Table 1 except the key predictor of interest, the average score so far. It suggests that modeling the effect as linear and symmetric for high and low average scores is reasonable.[4]

Effect size. The key point estimates from Table 1, B=-.1, imply that as the average score of previous applicants on a given day increases by one standard deviation (SD=.75), the expected score for the next applicant drops by about .075. To undo such a decrease, an applicant would need 30 more points on the GMAT, 23 more months of experience, or 0.23 more points (in a 1-5 scale) in the assessment of the written application.

Another benchmark for the effect size comes from the interview's subscores. Estimating a regression predicting the overall score with these five subscores as predictors, and all covariates from Table 1, we find that the aforementioned effect size of .075 is equivalent to the interviewee increasing her communication rating in the interview by about .33 SDs, or her interest in the school by .89 SDs.[5]

Heterogeneity. We consider heterogeneity across reviewers and days. For the former, we estimated the full specification (Column 4 in Table 1) for each interviewer separately. Out of the

---

compared to that in the real dataset of B=-.088 (nearly 1/50 the size). Thus, this concern of spurious correlations is not a problem in our dataset.

[3] We also estimated ordered probit and logit regressions. For column 4 the key point estimates are B=-155, p=.006 and B=-.250, p=.014, respectively. We restricted the sample to the third interview onwards believing that is where narrow bracketing is likely to have the most impact. Including all interviews Bs for columns 1-4 are -.089, -.079, -.076, and -.057, the first three have ps<.01, the latter has p<.05.

[4] For ease of exposition we truncated the x-axis at averages of 2 and 4 from above and below because very few observations (less than 5%) are in those 8 bins. Supplemental materials include a table with all values.

[5] Point estimates for the subscores when predicting overall score: communication (B=.286), driven (B=.262), team (B=.168), doer (B=.247), interest in school (B=.120). The SDs of these variables, controlling for observables from Table 1: .79, .71, .72, .71, .68. Impact of 1SD of communication on overall score, then, is .286*.79=.226. Dividing by .075, the impact of average score, we obtain 33%, which means that a one standard deviation increase in the average score so far has an impact equivalent to 33% of an increase in the SD of the communication score. Other calculations are analogous.

31 interviewers in the sample, 18 have enough interviews to allow such estimation given the large number of predictors. All but one of these 18 models result in a negative point estimate for the average score so far. Eight of the 17 negative estimates are significant at the 5% level. The single positive point estimate is not significant (p=.496). The aggregate pattern we observe, then, is not driven by a small subset of interviewers.

Two anonymous referees suggested examining the moderating role of variability in previous scores. The intuition is that after rating three candidates in a row with a 4, an interviewer may be more reluctant to give another 4, than after rating three candidates with a 3, 4, and 5 respectively. Consistent with this prediction, the effect is twice as large for ratings following a set of identical scores versus heterogeneous ones (B=-.111 vs. B=-.059). Despite the dramatic difference in point estimates, it is not statistically significant; we lack the power to detect sensibly sized effects.

Examining alternative mechanisms

We consider two alternative explanations for the previous findings. The first is a contrast effect. If interviewers employ recently seen applicants as a reference, then applicants following stronger applicants will seem weaker to interviewers, and those following weak ones will seem stronger, leading to a negative correlation in ratings within a day.

The second alternative explanation is non-random sequencing of applicants: if stronger candidates tend to be followed by weaker ones in the daily scheduling of interviews (or vice versa) then we also would expect a negative correlation of ratings within a day. We address each of these explanations next.

Contrast effects vs. narrow bracketing. We tested two sets of predictions that allow us to disentangle these alternative mechanisms. The first involves interview subscores. Recall that

each applicant is rated on five subscores concerning a specific attribute (e.g., communication skills) in addition to the holistic overall score analyzed so far. Because the evaluation of these specific attributes is more perception based and specific, one may expect them to be more susceptible to contrast effects; if our core finding were driven by a contrast effect, these subscores should also show an (arguably more pronounced) effect. For example, the contrast between an eloquent applicant and an inarticulate one seen back to back should be starker than that between applicants who differ in their overall strength aggregated across a broad range of attributes. Moreover, research on contrast effects in person perception shows such contrasts occur only through specific and relevant attributes (Higgins, Rholes, & Jones, 1977; Srull & Wyer, 1979), suggesting an overall contrast effect is likely to be the downstream consequence of specific attribute contrasts, and hence the latter a necessary condition for the former.

The *opposite* prediction follows from the narrow bracketing account. Because interviewers are unlikely to be concerned about keeping a balanced distribution of each subscore, and they may even have difficulty remembering the subscores they gave to previous applicants, subscores should more weakly, if at all, be influenced by previous subscores.

The second set of analyses involves the moderating role of how far into the set of daily evaluations an interviewer is. If interviewers engage in narrow bracketing, then as the day is about to end, imbalances should be particularly aversive, and the inclination to respond to the existing ratings stronger. The contrast effects literature on person perception is too nuanced to make an unambiguous prediction in this case. For example, depending on whether interviewers are focusing on similarity or differences among candidates, or whether previous candidates are extreme or moderate on a given attribute, one would expect contrast effects to get stronger or weaker, or even to become *assimilation* effects (for a review see Wheeler & Petty, 2001)

This second prediction is hence asymmetric (Larrick & Wu, 2007). If the impact of previous ratings does not increase as the day is about to end, the narrow bracketing account would be an inadequate explanation for the data. However, if the effect does get stronger toward the end of the day, a contrast effect mechanism would not be ruled out.

We conducted regressions analogous to those reported in Table 1 on the subscores. For example, we estimated the impact of the average communication skill score, given to previous applicants on a given day by a given interviewer on the communication score given to the current applicant (B=-.009, *p*=.748). The relationship was also weak and not significant for being driven (B=-.037, *p*=.279), ability to work in teams (B=-.027, *p*=.411), accomplishment (B=-.052, *p*=.073), and interest in the school (B=.025, *p*=.531). To reduce noise, we also averaged the five subscores and conducted the analyses on that average as if it were a sixth subscore (B=-.051, *p*=.223). Because subscores do not show the daily-horizons effect, we conclude that contrast effects are an unlikely explanation for our findings.

As noted by an anonymous referee, the greater specificity of the subscores may make them insufficiently ambiguous to exhibit biases. Providing a definite answer to this concern requires unavailable data: the relative ambiguity of subscores versus overall scores in the minds of the interviewers. The subscores in the dataset, however, seem similarly ambiguous to us as those used by scholars examining priming effects in person perception studies (e.g., assessing if Donald is kind or reckless, see Thompson, Roman, Moskowitz, Chaiken, & Bargh, 1994; Winter & Uleman, 1984).

We now move on to assessing the moderating role of having accumulated a larger number of interviews in a given day. We estimated regressions separately for subsets of interviews occurring in a particular serial position. The point estimates of interest, those of the

average score so far, are plotted in Figure 2. As predicted, the impact of previous scores grows stronger in size and significance as the day progresses.

Is applicants' objective strength negatively serially correlated within day? We estimated regressions predicting applicants' GMAT and job experience with the average scores of previous interviews. These are, in effect, placebo tests: if our interpretation of the data is correct, we do not expect to replicate our core finding in these specifications. Columns 5 and 6 of Table 1 show small, *positive*, and non-significant effects on these placebos. This evidence is hence inconsistent with candidates' objective strength, accounting for our core finding.

General Discussion

Building on the choice bracketing literature, which shows that decision makers insufficiently take into account the aggregate consequences of many similar decisions (Read et al., 1999), we examined narrow bracketing in judgment. In line with research showing that individuals put too much weight on an individual case and too little on background information (Brenner et al., 2005; Griffin & Tversky, 1992; Massey & Wu, 2005), we conjectured that people conducting sequences of subsets of judgments would insufficiently take into account other judgments they have already conducted or will conduct in the future. As a result, people avoid generating subsets of judgments that deviate much from the expected overall distribution.

We found support for this prediction using data from over 9000 interviews of MBA applicants, we conducted analyses that suggest the evidence is inconsistent with non-random scheduling of interviews and sequential contrasts effects.

While we have focused on well-defined daily subsets, a similar bias may occur when people conduct larger sets of evaluations and generate subsets spontaneously in their minds. Imagine, for example, a judge who must make dozens of judgments a day. Given that people

underestimate the presence of streaks in random sequences (Gilovich, Vallone, & Tversky, 1985), the judge may be disproportionately reluctant to evaluate 4, 5, or 6 people in a row in too similar a fashion, even though that "subset" was formed post-hoc.

We propose three mechanisms that may explain narrow bracketing in judgment. The first is based on the belief in the law of the small numbers (Tversky & Kahneman, 1971): upon giving a set of positive (negative) judgments, interviewers may form an expectation that a weaker (stronger) candidate "is due," or they may attempt to correct for perceived errors in their ratings based on deviations from expectations in their distribution. The second is that interviewers engage in mental accounting (Thaler, 1985, 1999), simplifying the task of maintaining a given long-term target of positive evaluations by applying their target to each "daily account." The third is that interviewers themselves do not engage in narrow bracketing but believe that people evaluating their performance do, and thus avoid unrepresentative subsets in an attempt to please their audience (as suggested by work on accountability; for a review, see Lerner & Tetlock, 1999).

These mechanisms are not mutually exclusive, and they may coexist. Future research could examine their relative importance in narrow bracketing and, perhaps more importantly, establish additional consequences of such psychological processes on everyday judgments.

References

Brenner, L., Griffin, D., & Koehler, D. J. (2005). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes, 97*(1), 64-81.

Gilovich, T, Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*, 295-314.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*(3), 411-435.

Higgins, T. E., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology, 13*(2), 141-154.

Larrick, R. P., & Wu, G. (2007). Claiming a large slice of a small pie: Asymmetric disconfirmation in negotiation. *Journal of Personality and Social Psychology, 93*(2), 212-233. doi: 10.1037/0022-3514.93.2.212

Lerner, J., & Tetlock, P. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255-275.

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under-and overreaction. *Management Science*, 932-947.

Read, D., Loewenstein, G. F., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty, 19*(1), 171-197.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Srull, T. K., & Wyer, R. S. (1979). Role of category accessibility in the interpretation of information about persons - some determinants and implications. *Journal of Personality and Social Psychology, 37*(10), 1660-1672. doi: 10.1037//0022-3514.37.10.1660

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science, 4*(3), 199.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making, 12*(3), 183-206.

Thompson, E. P., Roman, R. J., Moskowitz, G. B., Chaiken, S., & Bargh, J. A. (1994). Accuracy motivation attenuates covert priming: The systematic reprocessing of social information. *Journal of Personality and Social Psychology, 66*(3), 474.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127*(6), 797.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*(2), 237.

Tables

Table 1

Regressions predicting interview score (and GMAT and Job Experience) with the average of previous interview scores that day by the same interviewer

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	**Interview Score (1-5)**				***PLACEBOS*** **GMAT (250-800)**	***PLACEBOS*** **Experience (in months)**
Specification	Baseline	Interviewee controls	Interview controls	Score (1-5) of written application	Same as (3)	Same as (3)
Average interview score	**-0.116*** **	**-0.110*** **	**-0.105*** **	**-0.088** **	**0.085**	**0.251**
Given by same interviewer to previous interviewees that day (1-5)	***(0.038)***	***(0.035)***	***(0.036)***	***(0.035)***	***(2.063)***	***(0.959)***
GMAT score of applicant (/100)		0.244***	0.250***	0.079**	--	1.140**
		(0.036)	*(0.035)*	*(0.032)*	--	*(0.495)*
Months of job experience of applicant (/100)		0.324***	0.319***	0.254***	10.363**	--
		(0.057)	*(0.055)*	*(0.055)*	*(4.541)*	--
Number of interviews by same interviewer that day						
Total			-0.000	0.001	0.845	0.504*
			(0.012)	*(0.012)*	*(0.676)*	*(0.290)*
So far			-0.018	-0.010	-0.461	0.008
			(0.013)	*(0.014)*	*(1.275)*	*(0.360)*
Score given by *reader* of application				0.340***	24.190***	
				(0.044)	*(1.719)*	
Other controls						
Month*year dummies (k=12*9)	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Interviewer dummies (k= 21)	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Interviewee gender, race (k=9), age & age-squared	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Interview's time (k=12) & location (k=4)	*No*	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Number of observations	4,456	4,312	4,312	3,754	3,754	3,754
R-squared	.322	.381	.387	.484	.288	.726

Notes: Sample consists of interviews of MBA applicants between years 2000 and 2009 by 31 interviewers working for the admissions office of a private business school. Entries are pointe estimates from OLS regressions. Standard errors, clustered by interviewer, printed in parentheses. Sample is restricted to 3rd and later interviews for a given interviewer on a given day. GMAT and job experience (in months) are divided by 100 to arrive at readable point estimates. *k* indexes the degrees of freedom used by a variable (when greater than 1).
*,**,*** indicates significant at the 10%,5% and 1% level respectively.

Figure 1. Relationship between previous scores in the day by an interviewer, and the next score.

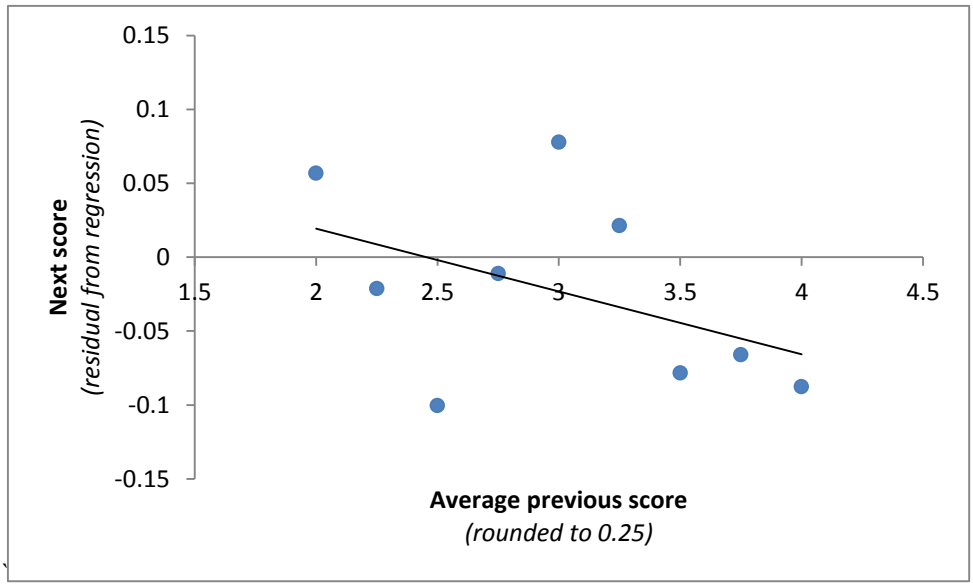
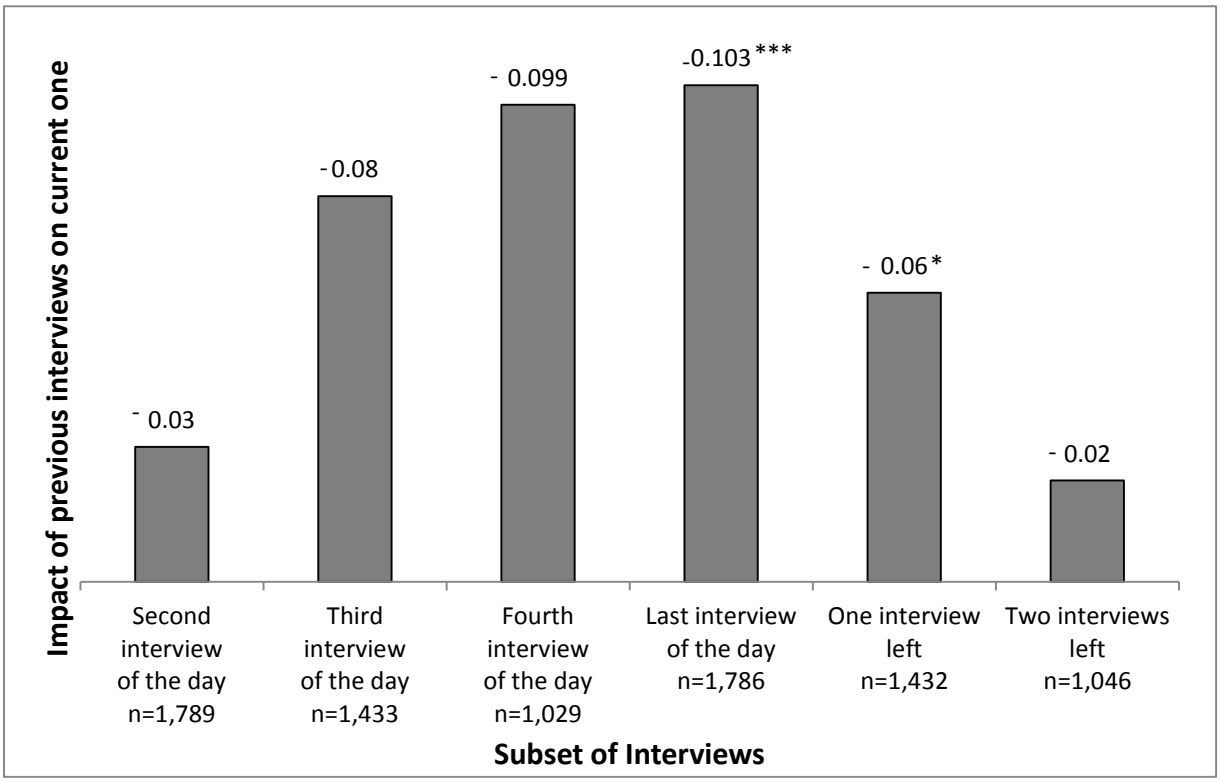


Figure 2. Impact of previous interviews is largest for last interview of the day.



Notes: Sample consists of interviews of MBA applicants between years 2000 and 2009 by 31 interviewers working for the admissions office of a private business school. Bars report pointe estimates from OLS regressions with the interview score (1-5) as the dependent variable and the average score of previous interviews that day as the predictor of interest. Regressions include all covariates employed in Column 4 of Table 1. The average number of interviews per day is 4.5 (SD=1.9). Sample sizes vary as a function of the number of interviews that fit criterion and that have non-missing information on any of the variables.
*,*** indicates significant at the 10% and 1% level respectively. *n* indicates the number of observations in the subset of interviews.