

Essays in Environmental and Labour Economics

Nikolai Cook

A thesis submitted in partial fulfilment of the requirements for the
Doctorate in Philosophy degree in Economics

Prepared on August 20, 2020

Department of Economics
Faculty of Social Studies
University of Ottawa

© Nikolai Cook, Ottawa, Canada, 2020

This document is a collection of articles that I have authored during the course of an Economics Doctoral Program at the University of Ottawa. I am grateful to my two co-supervisors, Anthony Heyes and Abel Brodeur, for countless hours of guidance and their exercise in limitless patience. Thanks are owed to my thesis committee members: Myra Mohnen, Matthew Webb, and Myra Yazbeck. Adam Lavecchia provided insightful comments for this research, as did my external examiner Casey Warman. Lastly, I am grateful to many others, far too numerous to name, for the time and effort they have generously given me.

The articles included here reflect my research in environmental and labor economics. They are thematically quite different but share a common thread of quantifying inputs of labor productivity. Outdoor temperatures, the race-gender pairing of employer and employee, and student aid were all found to affect how well people performed on tasks or how well they must be compensated.

The first chapter was written last and served as my job market paper. *Brain Freeze* presents evidence that cold temperatures have a detrimental effect on cognitive function. Identification comes from the quasi-random assignment of temperatures on exam days.

The second chapter was my first research paper. *A Boss Like Me* is a discrete choice experiment conducted on Mechanical Turk that quantifies the willingness to pay of employees for race and gender attributes of potential employers. Identification comes from randomizing which attributes are displayed to respondents, such as differing pay rates and employer race.

The third chapter, *Student Aid Increases Performance* presents evidence that student aid, in the form of the Ontario Tuition Grant, increases grades for those already enrolled in studies. Identification comes from a differences-in-differences design and panel fixed effects.

Each chapter contains its own relevant literature reviews. At the end of this document, a common bibliography is presented.

Contents

1	Brain Freeze	1
1.0.1	Abstract	1
1.0.2	Thanks	1
1.0.3	Ethics and Collaboration	1
1.1	Introduction	2
1.2	Literature: A selective review	4
1.2.1	Temperature (especially cold) and mental function	5
1.2.2	Adaptation	7
1.2.3	Projected change in cold	8
1.3	Data	9
1.4	Methods	11
1.5	Results	13
1.5.1	Basic plot	13
1.5.2	Linear	13
1.5.3	Non-linear	14
1.5.4	Heterogeneity	15
1.6	Cumulative effects	16
1.7	Adaptation	17
1.7.1	Organizational	17
1.7.2	Individual	19
1.7.3	Biological	22
1.8	Robustness	24
1.9	Conclusions	26
1.10	Tables and Figures	28
1.11	Appendices	46
2	A Boss Like Me	54
2.0.1	Abstract	54

2.0.2	Thanks	54
2.0.3	Ethics and Collaboration	55
2.1	Introduction	55
2.2	Methods	57
2.2.1	The Setting: Amazon’s Mechanical Turk	57
2.2.2	Job Offers	59
2.2.3	Data Manipulation	63
2.3	Results	64
2.3.1	Sample	64
2.3.2	Approach 1: Fixed Tasks	65
2.3.3	Approach 2: Aggregate Choice Patterns	67
2.3.4	Approach 3: Probabilistic Choice Modeling	68
2.4	Conclusions	75
2.5	Tables and Figures	77
2.6	Appendices	94
3	Student Aid	109
3.0.1	Abstract	109
3.0.2	Thanks	109
3.0.3	Ethics and Collaboration	110
3.1	Introduction	110
3.2	Context, the Grant, and Literature Review	111
3.2.1	Ontario Student Aid Circa 2011	111
3.2.2	The 30% Off Ontario Tuition Grant	112
3.2.3	Student Aid In Other Contexts	113
3.3	Data and Sample Restrictions	114
3.3.1	Summary Statistics	115
3.4	Identification and Econometric Models	116
3.5	Results	117
3.5.1	Graphical Evidence	118
3.5.2	Course Performance Increase With Aid	118
3.5.3	Academic Probation Decreases With Aid	120
3.5.4	Failures Decrease With Aid	120
3.5.5	STEM Students Benefit More From Aid	120
3.5.6	Wealthier Students Perform Better	121
3.5.7	Unsubsidized Behaviors Decrease With Aid	122
3.5.8	Student Aid Has A Larger Effect In Winter	122
3.5.9	Student Aid Decreases Graduation Rates	123

3.5.10	No Effect On Persistence Into Second Year	124
3.5.11	Robustness to Alternative Clustering and Standardization Strategies	124
3.6	Conclusion	126
3.7	Tables and Figures	127
3.8	Appendices	143

Chapter 1

Brain Freeze: Outdoor Cold and Indoor Cognitive Performance

1.0.1 Abstract

We present first evidence that outdoor cold temperatures negatively impact indoor cognitive performance. We use a within-subject design and a large-scale dataset of adults in an incentivized setting. The performance decrement is large despite the subjects working in a fully climate-controlled environment. Using secondary data, we find evidence of partial adaptation at the organizational, individual and biological levels. The results are interpreted in the context of climate models that observe and predict an increase in the frequency of very cold days in some locations (*e.g.* Chicago) and a decrease in others (*e.g.* Beijing).

1.0.2 Thanks

We wish to thank the Managing Editor and two anonymous reviewers for their helpful comments and suggestions. We are grateful to Sandeep Kapur, Soodeh Saberian, Abel Brodeur, Ian Mackenzie, Lata Gangadharan and participants at Monash University, LEEP and EAERE for helpful conversations. Errors are ours.

1.0.3 Ethics and Collaboration

This research was completed with administrative data accessed under University of Ottawa Research Ethics Board file number 11-17-15. It was completed

in collaboration with Professor Anthony Heyes. The student's contributions include but are not limited to conception of research question, statistical analysis, and drafting of chapter.

1.1 Introduction

How is the cognitive performance (“mental productivity”) of people working indoors, in climate-protected environments, impacted by outdoor cold? To what extent can adaptation at the organizational, personal, or biological level insulate against any decrement in performance?

This chapter provides what we believe to be first evidence that outdoor cold has a detrimental impact on performance, and to speak in detail to issues of adaptation. Data comes from a large sample of subjects in a fully-incentivized setting.

Understanding the link from exterior temperature to indoor work is a key step in any projection of how changing climate might impact productivity in sectors that are not as obviously climate-exposed as, for example, agriculture and tourism. While the attention of climate research in economics has been on increasing average temperatures and the effects of hot days on human outcomes, there is a dearth of evidence of any impacts of cold. This is an important gap in knowledge because climate models predict changes in the frequency of cold weather.¹ Even as average temperatures increase, some places will experience more very cold days by the end of this century (e.g. Chicago), while other places will experience less (e.g. Beijing). The effect of cold on the human body and behavior is distinct from that of heat and works through different channels. Furthermore, there exists evidence that the mechanisms for adaptation are different.

The outcome data that we use for performance is 638,238 exams taken by 66,715 adult students over a 9 year period at the University of Ottawa, a large, comprehensive, research-intensive public university. It operates from a main

¹Historically Chicago (with a mean December temperature of -3°C) has averaged 11 days in December where temperature remained below freezing for the whole day and a further 16 days in a typical January. The number of cold days in that and other mid-latitude North American cities such as Detroit and Toronto, is projected to increase between now and end of century due to arctic warming and increasing instability in the polar vortex [Kolstad et al., 2010, Cohen et al., 2018]. Beijing has a winter temperature profile similar to that of Chicago and is projected to get less cold days, particularly due to predicted changes in polar vortex states Kretschmer et al. [2018].

campus located in the heart of the capital city. While the extent to which impacts on exam performance would also be seen in workplace productivity is an open question, academic scoring reflects a clean measure of mental proficiency which, at a minimum, seems likely to correlate with performance in a range of brain-intensive work tasks. At least three features of our setting make it an ideal context to explore our research question:

(1) It provides good quality cognitive performance data on a large number of working age adults in an incentivized setting under *cold* and *very cold* exterior conditions (average daily temperature in our sample ranges from -17°C at the 5th percentile to 5°C at the 95th). The data's panel structure means we observe the same subject's performance under alternative outdoor-temperature treatments (on average around ten per subject), allowing inference based on within-subject variation. This expels any time invariant within month unobserved characteristics of individuals that might influence performance.

(2) The nature and scheduling of the cognitive tasks faced by subjects are determined far in advance and are insensitive to subsequent temperature realizations. This allows us to rule out selection effects due to displacement-in-time of activity in response to conditions that could contaminate inference in other settings.

(3) While outdoor temperatures vary widely, we are able to provide direct evidence that the indoor temperature for subjects are held almost exactly constant by modern climate-control technology. As such, the most obvious technological protection against extreme temperature is fully-exploited, and any effects we identify account for that margin of adjustment.

Secondary data allows us to investigate non-organizational adaptation. While an employer, for example, can heat the workplace, there are actions that individuals can take to protect against outdoor temperature conditions. We test whether reducing direct exposure through living close to place of work provides mitigation. To investigate the hypothesis that personal protection against extreme cold can be purchased (buying better winter clothing, using taxis on cold days, etc.) we investigate how temperature sensitivity relates to a proxy for subject income. To probe biological adaptation to cold conditions we (a) compare the sensitivity to treatment of domestic students with those from overseas (in particular from a set of hot countries) and, (b) examine how the sensitivity of the latter group evolves with repeated exposure.

We find a negative impact of outdoor temperature on indoor performance. The effect is substantial. In our preferred specification, which includes student fixed effects, year fixed effects, and controls for other weather conditions, a ten

degree (1.75 standard deviations) Celsius colder outdoor temperature on exam day causes a reduction of about one-twelfth (8.09%) of a standard deviation in performance. The magnitude and significance of the effects prove highly robust to a wide range of tests. We speak to issues of mechanisms indirectly by characterizing the (less-than-complete) efficacy of adaptive strategies at various levels. While our study relates to adults taking university-level exams, such performance effects might be expected in a wider range of mentally-demanding tasks in the workplace.

The rest of the chapter is organized as follows. In Section 2, we review some pertinent existing research. In Section 3, we detail our administrative and weather data. Section 4 presents our identification strategy. Section 5 details our main results. Section 6 explores cumulative effects of cold. Section 7 details results on adaptation. In Section 8, we challenge the robustness of our results. Section 9 concludes.

1.2 Literature: A selective review

Temperature is increasingly recognized as an important factor in many outcomes of interest to economists. The effect of temperature realizations on productivity have been characterized at the economy level by Dell et al. [2012], United States county level by Deryugina and Hsiang [2014] and plant-level by Zhang et al. [2018]. Recent papers have found effects of hot weather on human outcomes including morbidity [Bleakley, 2010, Schwartz et al., 2004], mortality [Barreca et al., 2016, Burgess et al., 2017], productivity [Somanathan et al., 2015] and decision-making [Heyes and Saberian, 2019]. In such studies, the temperature observations have typically fallen in the range above 25°C, implying little or no power to uncover impacts of low temperatures.²

²Lee et al. [2014] regress outdoor temperature on speed of completion of a routine clerical task by bank employees in Tokyo. They find a negative and significant coefficient on their quadratic temperature term, consistent with a *positive* impact of either extreme heat or extreme cold on productivity. However; (1) The mean and standard deviation of outdoor temperature in the table of summary statistics are 17°C and 5°C respectively, suggesting few observations in the temperature range of interest to us. (2) The authors do not allow for the possibility of asymmetric impacts of heat versus cold by (for example) applying non-parametric methods.

1.2.1 Temperature (especially cold) and mental function

Among research linking outdoor temperature to cognitive performance, such as Graff Zivin et al. [2018], find that short-run changes in temperature negatively impact the cognitive performance of children above 26°C but find little evidence of longer-run effects.³ Park [2016] studies children taking standardized exams in a panel of New York City schools during the month of June. He finds that performance is compromised by 0.22% per 1°F (0.55°C) rise above 72 F (22.2°C). Goodman et al. [Forthcoming] focus on longer run effects of hot weather across the school-year, finding that each 1°F increase in school year temperature reduces the amount learned that year in U.S. schools by about 1%.

Zivin et al. [2018] use data from the fixed date of the National Chinese Entrance Exam to estimate the effects of outdoor temperature on cognitive performance. They find that, in a setting without air conditioning or the ability of students to sort by location, a 1°C increase in summer temperatures (mean of 23.2°C) reduced performance by 0.029 standard deviations.

Research on the effects of cold temperature on mental performance and productivity is less developed. With one notable exception, the evidence that does exist relates exclusively to *contemporaneous* temperature. In other words performance and behavior *during* exposure. Pilcher et al. [2002] provides a meta-analysis and Taylor et al. [2016] a survey.

Without identifying a mechanism, various experimental studies have shown that contemporaneous exposure in the range - 20°C to 10°C can reduce memory function [Thomas et al., 1989, Patil et al., 1995], consistency of decision making [Watkins et al., 2014], and speed in pattern recognition and number comparison [Banderet et al., 1986]. Studying driving behavior in cold conditions, Daanen et al. [2003] note that cold can impair mental function and thus increase accidents, observing a 16% decrement in performance of drivers in simulated conditions at 5°C compared to 20°C.

There are several channels that might link cold to compromised cognitive performance. In their survey, Cheung et al. [2016] emphasizes the depleting effect of thermoregulation. The initial response to short-term cold exposure

³They explicitly acknowledge that they can speak to high temperatures only: “Since these tests were predominantly given during the warmer periods of the year, our analysis of short-run temperature effects will only be informative for temperatures in this range” [Graff Zivin et al., 2018, p.84]. In their dataset, for example, the mean temperature on day of test is 22.5°C and standard deviation 4.9.

is cutaneous vasoconstriction, reducing blood flow to the skin and extremities. This serves to decrease the thermal gradient between the body and environment. While this is effective in maintaining body core temperature, it simultaneously causes discomfort. As exposure persists, heat maintenance requires the depletion of limited carbohydrate stores [Bell et al., 1992] which has been shown to decrease manual dexterity, motor coordination, work tolerance, and “perceptual discomfort that can effect cognition” [Cheung et al., 2016, p.155]. Exposure to cold conditions also alters the concentration of central catecholamines in humans which has been linked to “... a detrimental effect on cognition as brain regions such as the prefrontal cortex are reliant on these neurotransmitters for normal function, ... (as such) there is a plethora of evidence which demonstrates that tyrosine supplementation improves cognitive function during acute cold stress” [Taylor et al., 2016, p.372]. Breathing very cold air can also irritate the human respiratory system, potentially damaging mood [Hartung et al., 1980], while even brief cold exposure can elevate hormonal stress markers [LaVoy et al., 2011].

A parallel body of research highlights the role of psychological mechanisms. Consistent with the classic “distraction theory” of Teichner [1958], cold conditions may provide alternative stimuli and thus interrupt focus which would otherwise be applied to the cognitive task in hand (“*i.e.*, attention is focused on feeling cold rather than competing the cognitive task provided” [Taylor et al., 2016, p.372]). Uncomfortable temperatures might also influence motivation and performance via their negative effect on mood or sentiment (see citations in Noelke et al. [2016]). The case for the importance of psychology is reinforced by studies such as Rai et al. [2017], which show that the attitudes and behaviors of experimental subjects can even be influenced by temperature *cues*, such as photographs of cold places.

While such studies are suggestive, they offer little help in understanding what the wider impact of cold outdoor temperature might be across the economy, since the vast majority of mentally-taxing work in cold countries is done indoors. Indeed, in most industrialized countries the median adult spends more than 90% of their time indoors, particularly during cold weather [Nguyen et al., 2014]. ? finds similar effects for children, as when especially cold weather occurs more time is spent inside.

To our knowledge, the only study examining the sustained impairment due to cold exposure after stimuli is removed is Muller et al. [2012]. They track a sample of 10 young adults during *and after* being cooled in a temperature-controlled chamber at 10°C. Working memory, choice reaction time and exec-

utive function declined during exposure, and impairments sustained an hour after exposure. This points to the possibility of the impact of exposure to outdoor cold being something that the subject imports when they move indoors. Relatedly, Heyes and Saberian [2019] argue that uncomfortable outdoor temperature might affect indoor performance *even if the subject is not directly exposed to it*. For example, extreme cold may prevent or discourage subjects from going outside to ‘stretch their legs’. Lack of fresh air has been linked experimentally to outcomes such as decreased mental function [Chen and Schwartz, 2009] and depressive mood [Cunningham, 1979].

1.2.2 Adaptation

Adaptation to cold outdoor temperatures might occur at various levels (for example national, municipal, organizational, individual) and over time. In this chapter, we present short-run analyses that will net out avoidance measures that are based on historical climate, such as locational sorting, technology adoption and building design.

The first and most obvious short-run protection against cold weather is to move indoors. The extent of protection afforded by a building plausibly depends on the effectiveness of its interior heating. At the other end of the temperature spectrum, the analogous protective benefits of air conditioning have been explored in a number of studies. Park [2016] study New York City children taking Summer exams, and does not find a significant protective benefit to air conditioning. He does note that of schools with air conditioning installed, up to 40% were deemed defective by an independent survey. In contrast, Goodman et al. [Forthcoming] finds that school level air conditioning offsets most of the potential learning decrement due to heat.⁴

A related literature studies the mitigative effects of other ‘technologies’, such as investment in high quality winter clothing [Mäkinen, 2007]. We will explore pecuniary channels of self-protection later.

Biological adaptation may also be physiological or psychological, though evidence on each is comparatively scarce. Teichner [1958] developed the concept of *psychological cold tolerance* “... which was conceived as depending largely on the individual’s familiarity with cold and on his anxiety level. These

⁴Goodman et al. [Forthcoming] uses a triple-difference strategy combining within-student observations with within-school variation status in cooling status over time. The only threat to such an approach is the possibility that the timing of A/C installation was correlated with other unobserved improvements in learning environment.

are factors reflected in the individual's subjective reactions which should not be ignored when discussing performance in the cold." [Enander, 1984, p.370]. In terms of such habituation there is some evidence of changes in attitude to cold after repeated exposure. In early work, Fine [1961] showed that subjects evaluate 'cold' less on a cold-warm scale after repeated exposure. Enander et al. [1980] compared the response to cold of subjects accustomed to working in cold conditions (meat cutters) against office workers. While there was no difference in physiological response, they found evidence consistent with psychological adaptation. The accustomed group experienced significantly less cold sensation and pain than the unaccustomed group. Another study consistent with physiological adaptation is Tochiara [2005], who found that the rectal temperatures of a sample of coldstore workers fell less when exposed to a temperature of -20°C for 60 minutes than did those of the control sample.⁵ Several studies have found evidence consistent with increased brown adipose tissue ('brown fat') among those exposed to frequent cold (for example Blondin et al. [2014]).

Overall, the bulk of the evidence points to a primarily psychological adaptive process to cold. This provides an interesting contrast to the analogous evidence on adaptation to heat exposure. "(T)he evidence of physiological adaptations from longitudinal cold exposure is equivocal [Launay and Savourey, 2009], while the dominant adaptation is a perceptual habituation and desensitization to cold stress rather than large-scale systemic physiological changes of the sort seen with heat acclimatization" [Cheung et al., 2016, p.155].⁶

1.2.3 Projected change in cold

It is commonly assumed that as climate warms, the distribution of daily temperatures will see a rightward shift towards warmer averages. In isolation, this

⁵Brazaitis et al. [2014] immersed 10 male subjects in 14°C water and timed how long it took for body temperature to drop to 35°C . On day 1 the average cooling time was 130 minutes, on day 14 cooling time had fallen to 80 minutes. The authors suggest a reduction of temperature gradient as a possible adaptation to cold.

⁶The abstract in the survey of physiological adaptation by Daanen and Van Marken Lichtenbelt [2016] ends: "Dedicated studies show that repeated whole body exposure of individual volunteers, mainly Caucasians, to severe cold results in reduced sensation but no major physiological changes. ... (H)uman cold adaptation in the form of increased metabolism and insulation seems to have occurred during recent evolution in populations, but cannot be developed during a lifetime in cold conditions. Therefore we mainly depend on our behavioral skills to live in and survive the cold" [Daanen and Van Marken Lichtenbelt, 2016, p.104].

would indicate that problems of extreme cold temperatures may be alleviated due to warming temperatures. However, while this turn out to be the case in many places - in which case the effects that we uncover in the chapter will deliver a previously unaccounted for benefit of climate change - in others it will not.

Hansen et al. [2012] showed that the chances of unusually cool seasons have risen in the past 30 years, coinciding with the observed rapid global warming. One mechanism through which this has been studied is a weakening of the polar vortex, which makes easier the periodic southerly movement of cold Arctic air masses. Kolstad et al. [2010] and Kretschmer et al. [2018] show that in the past several decades the frequency of weak polar vortex states has increased, which has been accompanied by subsequent cold extremes in the mid-latitudes, including North America, Europe and northern Asia. Kim et al. [2014] find evidence linking weakening of the vortex to Arctic sea-ice loss, consistent with the trends associated with climate change. “A handful of studies offer compelling evidence that the stratospheric polar vortex is changing, and that this can explain bouts of unusually cold winter weather (in North America)” (Francis, 2019).

1.3 Data

We obtained administrative data from the university as the basis for our measure cognitive performance. In particular, we observe the universe of grades achieved by undergraduate students for over 1.2 million courses. Our sample includes students who first enrolled for a course at the university in or after the Fall semester of 2007, and the latest courses we observe are those examined in December 2015. We connect this dataset with institutionally provided student information such as gender, age and address. Data on financial status by six-digit postal code comes from the 2016 Canadian Census of Population.

The academic year is split into two semesters. Fall-semester courses are taught from September through November, with final exams written in December. Because of our interest in cold we use these grades ($N = 638,238$) and the students that achieved them ($N = 66,715$) as the basis for our analysis.

That course-level grade is our dependent variable introduces a complication. While we hypothesize that exam day temperature impacts performance in the final exam, assessment for each course is based only partially on final exam performance. Other elements such as midterms or coursework completed during the semester also contribute. Academic regulations require that final

exam weight be no lower than 40% and no higher than 60%. The variation in weighting adds measurement error to the dependent variable which is uncorrelated with our regressor of interest.⁷ While such measurement error does not bias OLS estimates, it increases the associated standard errors making significance claims conservative. It also requires that in interpreting effect sizes, we use a multiplier to reflect that any impact of exam-day temperature on exam performance has a dampened impact on course-level performance. In our main specifications we impute the variation in exam performance as a factor of two times the variation in course performance, consistent with the assumption that the final exam carried 50% of the weight in every course. In doing so, a 5% decrement in overall course score maps to a 10% decrement in final exam score.

Daily meteorological data comes from the nearest Environment Canada weather station that provided consistent data across our period (Station ID 6105978) located 5.1 km from the centre of the campus. There is wide variation in the outdoor temperatures experienced by students on exam days, illustrated in Figure 1.

Summary statistics relating to course performance, student characteristics and weather are in Table 1. The average course grade is 71.98%, corresponding to a 'B' in the university grading scheme. Grades vary considerably within-student, the standard deviation is 10.31%, or two letter grades around the mean. Exam days are cold, averaging -5.13°C . Temperatures also vary considerably within-student, as a one standard deviation colder temperature is -10.81°C while a one standard deviation milder temperature exam day is above freezing. There is often snow falling (the equivalent of 2.12 cm)⁸ and snow already on the ground (2.46 cm). Female students account for 60% of the data while foreign students contribute 7.43%. We use a total of 638,238 exams, written by 66,715 students. The succeeding columns present summary statistics by gender and foreign status.

⁷The granularity of course grade reporting is an additional source of measurement error. Final course grades are recorded as letters, which correspond to a score interval. For example, an 'A' corresponds to a score in the interval 85-89%, which we then assign to the midpoint of its interval.

⁸Environment Canada uses a 10-to-1 conversion of water equivalent precipitation and snowfall.

1.4 Methods

In this section we detail the identification strategy used to estimate the causal impact of outdoor exam temperatures on indoor cognitive performance (imputed exam score).

Identification comes from quasi-random assignment of exterior temperatures to exam days. Fall semester exams are held in an exam period that runs from early in December until the university closes for the Christmas recess. The earliest and latest dates on which we observe exams in our sample are December 4 and December 21. Exams are held in one of three time slots (beginning at 9:30 am, 2 pm and 7 pm).⁹ The university releases the exam schedule in mid-October, much later in the semester than the final class enrolment deadline (mid-September).

Our results use a student fixed effects model estimated by Ordinary Least Squares (see, for example, Ebenstein et al. [2016]). Our main specification is:

$$Grade_{i,t} = \beta_0 + \beta_1 * Temperature_t + \Delta_t + \gamma_i + \eta_y + \epsilon_{i,t} \quad (1.1)$$

Where $Grade_{i,t}$ is the imputed exam performance for individual i taking a course where the final exam took place on day t . Our parameter of interest is β_1 , the coefficient of mean outdoor temperature on the date of exam. We explore the robustness of our estimate using alternative temperature measures later. The standard errors are clustered at the student level. Later, we demonstrate that results are robust to a number of other plausible clustering strategies.

The inclusion of student (γ_i) and year (η_y) fixed effects implies that identification comes from within-student and within-year variation. In other words, variations in the performance of individual subjects under alternative temperature treatments, within an exam period. Year fixed effects capture changes to course grades between years that are common across students including, for example, grade inflation.

⁹We do not observe students allowed to defer an exam to a date other than that mandated for the course, typically about 4% of the total. Deferment for reasons unrelated to temperature (family bereavement, religious holiday, etc.) are of no concern. Insofar as some deferments result from low exam-day temperature it is plausible that it works against the direction of any effect that we find, since postponement from a day that is unusually cold is likely to be to a later date that is less cold. However this is a valid caveat to hold in mind. Note that the university as a whole never closed on a regular business day or canceled an exam for weather-related reasons during the study period.

Δ is a vector of exam-day controls – precipitation on exam day and its interaction with temperature, relative humidity, snow on ground, windchill, day of week indicator variables and the date-in-month.

Inclusion of the interaction term between temperature and precipitation in our specifications reflects the common observation that damp cold may have a different effect than dry cold. For the same reason relative humidity is included as an additional control in our preferred specification.¹⁰ The interpretation of β_1 is the effect of a 1°C change in exam temperature on a dry day. There is zero precipitation on 45% of the days in our sample, and less than one millimeter of precipitation on 62% (see Figure A2 for a full distribution). A robustness exercise shows that effect sizes sustain even when we estimate on dry days alone. We also present estimates without precipitation, or its interaction, in an appendix.

Precipitation in December almost always means snow at this location. In addition to precipitation actually falling on a particular day, we also include accumulated snow on ground (measured by Environment Canada’s acoustic sensors such as the SR-50A). Accumulated snow might effect ease of travel, although it is worth noting that the municipal government exerts considerable efforts to the clearance of snow from sidewalks and streets in the city, as does the university on its campus. Actual experience of snow under-foot in the vicinity of a downtown location such as the university campus is likely quite different to conditions at the weather station.

Day-of-week fixed effects capture the possible effects of exam timing while date-in-month (as a continuous variable) captures any variation in exam performance correlating to *when* in the month an exam takes place. For example, including date-in-month helps if “difficult” courses tend to have exams scheduled later in the month, or if proximity to the holidays has an effect on exam performance.

In a supplementary analysis we explore the possibility of a non-linear relationship between outdoor temperature and indoor performance. To do this we estimate two models. First, our continuous temperature regressor in Equation 1 is replaced by a series of indicator variables corresponding to bins of width 2.5°C. Second, we use a series of indicator variables that organize temperature treatments into deciles.

¹⁰We report the estimates of precipitation, temperature \times precipitation and the other controls in Table A1.

1.5 Results

1.5.1 Basic plot

Figure 2 provides a simple plot of exam day temperature and exam performance, after adjusting only for year of exam. The size of markers is proportional to the number of observations in each 0.5°C bin.

Visual inspection suggests a positive association between performance and exam day temperature. We formalize this by plotting the line of best fit estimated by OLS with only year fixed effects.

While the absence of plausibly important controls means that such a plot and associated fitted line should be treated with caution, these initial effect sizes are substantial and prove robust to the inclusion of controls and their associated alteration of the temperature coefficient's interpretation.

1.5.2 Linear

Our main results are reported in Table 2. The dependent variable is expressed in hundredths of a standard deviation of exam score. Standardization of grades is across all years and students.¹¹

Column 1 presents our sparsest specification, containing student and year fixed effects and accounts for precipitation and the precipitation \times temperature interaction.¹² Column 2 adds controls for day-of-week. Column 3 controls for date-in-month. Column 4 through 6 add relative humidity, accumulated snow on ground and windchill, respectively.

In each column, the estimated coefficient on temperature is positive and statistically significant beyond the 1% threshold. Coefficient values are also stable across specifications. Column 6 presents our preferred specification, corresponding to Equation 1.

The coefficient on temperature is 0.809^{***} , suggesting that for every 1°C increase in exam day temperature, performance increases by 0.00809 standard deviations.¹³ The 90th and 10th percentiles of the temperature distribution

¹¹In Table A5 we standardize by year and course to find similar estimates.

¹²In Table A1 we also report our analysis without precipitation or its interaction with temperature. We then report the coefficient of precipitation and its interaction with temperature, and find both are negative and statistically significant. A specification in which we drop *all controls* is reported as a robustness exercise in Table 11, and delivers a main coefficient of 1.526^{***} .

¹³It is possible that exam markers adjust their grading standards in response to the quality

in the sample are 2.2°C and -14.7°C respectively. Hypothetically moving from a day at the 90th percentile in terms of temperature, to a day at the 10th percentile, delivers a decrease in temperature of 16.9°C . According to our preferred estimate this causes a substantial decrement in exam performance of 0.14 about one-seventh of a standard deviation. Equivalently, to deliver a reduction in performance of 0.1 or one-tenth of a standard deviation would require a 12.4°C decrease in outdoor temperature.

1.5.3 Non-linear

In Table 3 we repeat the exercise just described but replace the continuous measure of exam day temperature on the right-hand side of Equation 1 with a series of eight indicator variables. Each takes the value 1 if average temperature on exam day t fell in the range that defines the associated indicator's bin. Bins are constructed to be 2.5°C in width, built out from zero. The bin containing days with temperature below -15°C is the reference (omitted) category.

Each column in Table 3 replicates the combination of controls in the same-numbered column in Table 2. The preferred specification is again reported in column 6. The coefficients for each bin are broadly consistent across columns, suggesting that estimated non-linear effects are also robust to the inclusion of alternative control sets. The coefficients and associated 5% confidence intervals from the sparsest (column 1, left panel) and preferred specification (column 6, right panel) are plotted in Figure 3.

Figure 3 shows a negative impact of cold outdoor temperature on performance, which is roughly linear over the range that we study. The vertical axis scale in both figures is hundredths of a standard deviation. For example, in the right-hand panel of Figure 3, moving from a day in the 0°C bin to the -15°C bin reduces course grade by about 12% of a standard deviation.

While the overall trend seems to be roughly linear, here we note two interesting artifacts of Figure 3. The first is that the -15°C to -12.5°C temperature bin has an estimated effect that is worse than the colder temperatures below -15°C . We are relieved that when the data is divided in another reasonable manner (into deciles in Table A2 and Figure A1) we find results broadly con-

of responses in a particular pile of scripts. Insofar as that is the case it seems likely that the correlation between grading stringency and response quality is positive (the marker would apply laxer standards if she found the students performing poorly). This would imply that our estimated coefficient would understate the true effect size, making inference conservative.

sistent with those in Figure 3 while removing this anomalous negative effect for that temperature range. Second, exams with temperatures above zero seem to have disproportionately better results, suggesting we could enrich our specification with a kink. In Table 11 we winsorize our temperatures beyond the 0°C mark and find no meaningful differences to our main estimates.

1.5.4 Heterogeneity

In this subsection we investigate heterogeneity of effect size by sex, ability, and foreign status of the student. To do this we add to the preferred specification, in separate exercises, interaction terms between temperature and an indicator variable for the subsample in question. The results of these exercises are reported in Table 4.

In column 1 we interact temperature with an indicator that takes value 1 if the student is female. The estimated coefficient of 0.927*** is for a male student. The negative and significant interaction term implies that *ceteris paribus* female students are about twenty percent less sensitive to cold, consistent with research that has found women wear both more layers and more articles of clothing in cold weather, regardless of activity [Donaldson et al., 2001].

In column 2 we conduct the same exercise but with an indicator that takes value 1 if a student arrived at the university with an A (or 80) admission average. This applies to 43% of our sample of exams. The coefficient on the interaction term is large in value, -0.311***. The central estimate suggests that these high-admission students are roughly one third less cold-sensitive than their counterparts. This is not surprising given that most domestic (Canadian) students that admit as high achieving have already demonstrated an ability to perform well in winter examinations under comparable outdoor conditions in the context of their secondary school education, prior to attending university.

In column 3 we conduct the same exercise on foreign students, using domestic students as a baseline. Classification as foreign student is derived from paying international student fees to attend the university, or through immigration status. Perhaps unsurprisingly, foreign students are around 60% more sensitive to cold than domestic students. Almost all foreign students come from countries that are substantially less cold than Canada, and so are unlikely to be accustomed with such temperatures. We provide evidence of habituation or biological adaptation by investigating the performance of foreign students, both on arrival and through time, later.

1.6 Cumulative effects

While not our main focus, before turning to adaptation we investigate effects of temperature not just on the exam day, but during the preceding teaching semester.¹⁴

To do this we add to our preferred specification, a proxy of the total ‘cold’ experienced in the 30, 60 and 90 days prior to the exam. The measure that we use for cumulative cold is total heating degree days (HDD) over the period in question. A HDD is the number of degrees that the average temperature on a particular day is below 18°C, and is the standard measure used to quantify cumulative demand for heating in buildings. For example, if in a 30 day window half the days have an average temperature of 12°C while the other half have an average temperature of 17°C, the total HDD count over that 30 day window would be $(15 \times 6) + (15 \times 1) = 105$.

Table 5 reports the results of these three exercises. Columns 2, 3 and 4 include the total HDDs in the 30, 60 and 90 days prior to first exam, respectively.¹⁵

The results in this table are interesting for two reasons.

First, as a robustness check on our main result. The coefficient on our primary independent variable of interest, same-day temperature, is stable across columns. This suggests that we have isolated short from longer-run temperature effects. A potential challenge to our main specification is that temperature on exam day may be correlated with how warm or cold it had been in the lead up to the exam, such that failing to control for the latter would bias (or completely explain) our central estimates. Comparison of the columns in this table discourages the view that any such bias has substantially distorted our results. To ensure that this is not an artifact of the HDD measure, we report the results of analogous exercises using either average temperatures or much shorter pre-exam windows in Appendix Tables 2 and 3. We find our coefficient of interest is little-disturbed.

Second, in each of columns 2 through 4 the estimated coefficient on the pre-exam history of HDD is statistically significant. Temperature during the semester appears to have a significant impact on how students perform. However the sign is *positive*, implying cooler temperatures across the teaching term

¹⁴Evidence of the cumulative effect of temperatures on cognitive performance is mixed. For example, with respect to much warmer temperatures Goodman et al. [Forthcoming] found no cumulative effect of temperature on learning in United States schools with A/C.

¹⁵In Table A3 we use average temperatures leading to exam day, the results are similar.

are associated with improved performance. This is consistent with previous literature that finds unappealing outdoor temperatures can encourage substitution from outdoor leisure to indoor ‘work’ [Graff Zivin and Neidell, 2014]. For example, in column 2, if each day in the 30 leading up to the exam were one degree warmer, that would roughly offset exam-day temperature being one degree colder.

Another consideration could be cold temperatures leading to student sickness. While we do not have case-level data of, for example, admissions to the university clinic, we do analyze how short run temperatures leading up to the exam affect performance in Table A3. We find previous 1,3, and 5 day average temperatures leading to the exam have mixed signs and statistical significance. We note that this measure is imperfect and see examining the relationship between cold and sickness as a possible avenue for future research.

1.7 Adaptation

Central to any analysis of the costs of climate change is understanding the efficacy of adaptation. Analyzing adaptation also speaks indirectly to mechanisms that might underpin the effect that we have identified. We explore adaptation at three different levels.

1.7.1 Organizational

There are two temperatures that might influence how a worker performs, namely indoor and outdoor. The employer can control the former, but not the latter.

There are two separate questions that research in this area can address. First, to what extent is the technology of climate control effective in decoupling indoor from outdoor temperature. Second, insofar as it does lead to full or partial decoupling, to what extent does that mitigate the causal effect of outdoor temperature on the outcome variable of interest.

With respect to hot temperatures, recent studies provide evidence of only partial mitigation by air-conditioning. These share two important limitations. (1) Installation and quality of air-conditioning is unlikely to be randomly-assigned, and in many settings is plausibly correlated with unobserved characteristics (such as financial circumstances) of the school, business or other organization that might impact effect size through other channels. (2) To our

knowledge, the actual efficacy of the cooling technology is unknown.¹⁶

Winter heating in Ottawa public buildings is good, perhaps not surprising given that very cold temperatures are common. Employers in Ontario (including universities) are obliged by law to maintain a workplace temperature above 18°C. In light of this, internal temperatures experienced by our subjects are plausibly uncorrelated with outdoor temperature by design. However we tested this directly by working with campus building managers to measure and collect data on daytime interior temperature. The sample was collected during December 2018 for the 28 most important exam rooms by contribution to sample. Matching with outdoor temperature on the same day, we investigate the links between indoor versus outdoor temperature in exam rooms.

The data collected for Montpetit Hall Room 021 (MNT021) is presented in Figure 4. This is the largest room by contribution to sample, contributing 66,888 of the 638,238 observations that we use in our regressions. There are two important features of this plot. First, there is little variation in indoor temperature, fluctuating between $21.5 \pm 0.3^\circ\text{C}$ (reference lines at $\pm 1^\circ\text{C}$ of the room average are provided). Second, such variation as does exist does not look to be meaningfully correlated with outdoor temperature.

Figure 5 presents analogous diagrams for each of the 28 rooms (MNT021 is third from the left, second row). In each case we superimpose horizontal reference lines at the room’s average temperature $\pm 1^\circ\text{C}$. The figure tells us that all exam rooms are not equal in terms of the consistency with which internal temperature is maintained. In some rooms internal temperature fluctuates outside the $\pm 1^\circ\text{C}$ corridor, though even in these ‘leaky rooms’ there is little suggestion of correlation between outdoor temperature and what is going on outside.

We conduct two further exercises to test whether our central results are driven by imperfect climate control.

¹⁶Quinn et al. [2014] and Tamerius et al. [2013] present survey evidence on the relationship between indoor and outdoor temperatures in a sample of 327 buildings in New York City. For outdoor temperature ranges above 15°C they find a correlation between outdoor and indoor temperature to be 0.64 [Tamerius et al., 2013, Fig.1] despite air-conditioning penetration in that city at time of sample being 87.5%. Interesting given our focus is that for temperatures below 15°C the correlation coefficient between indoor and outdoor temperature is just 0.04. In general, heating space is easier than cooling it. In addition, modern air-conditioners are characterized by a ‘temperature drop’ - the maximum by which the refrigerant coils can reduce incoming to outgoing temperature - which for most common designs is less than 20°C. Even if working to its full potential, this places a bound on how cool the air-conditioned space can be kept when outdoor temperatures are very high.

First, we test the role of building age. Our sample includes both new and old buildings. For example, Tabaret Hall (TBT) was constructed in 1856. While spaces are well maintained, there is a concern that our results are driven by older buildings that do not meet modern standards. To explore this we divide buildings into two categories, ‘New’ (those completed after the year 2000), and ‘Old’ (the rest). This roughly splits our sample in half. Column 1 of Table 6 reports the results of adding to our main regression an interaction term that between exam-day temperature and an indicator variable that takes the value 1 if the exam room is located in a new building. The interaction term is negative, and marginally significant, consistent with our concerns. The estimated coefficient on temperature (0.837***) is now interpreted as the effect of temperature on performance for exams written in an old space. Writing in a new building is estimated to offset about 14% of the outdoor temperature effect.¹⁷

Second, we exploit the room temperature measurements reported in Figure 5 directly. Even within a building some rooms may be better temperature-controlled than others. In column 3 of Table 6 we report the results from running the specification from column 2 but excluding the exams taken in rooms identified as ‘leaky’ in Figure 5 (that is, those with temperature observations outside the $\pm 1^\circ\text{C}$ band). Under this restriction the coefficient of the new building \times temperature interaction term becomes much smaller and far from statistically significant at conventional levels.

Taken together, the evidence in this subsection supports our conjecture that the most obvious technological adaptation that an organization can use to protect employees against cold, namely climate control, is relatively fully-exploited. As such, the effects that we identify should be understood as already accounting for that base margin of protection.

1.7.2 Individual

Individuals plausibly have ways in which they might protect themselves privately from cold. We explore two. One approach is to reduce exposure by reducing commuting time. Another is spending on personal protection.

First, we examine the extent to which our effect dissipates with proximity

¹⁷For completeness we repeat the specification in column 1 but including course level fixed effects, as there may be a relationship between building age and course level. This is reported in column 2 in Table 6. The additional inclusion does not change results, and increases the statistical significance of the new building and temperature interaction term.

to campus. We note that residential location and commuting time is not randomly assigned in our setting. Students might reasonably be assumed to take account of climate when deciding where within the city to live, and results in this section need to be interpreted with that in mind. We add to the preferred specification a control for distance between campus and term address as recorded in the student record ('Distance'). We then linearly interact distance with exam day temperature. For completeness, we also add the interactions between distance and precipitation, and between distance and accumulated snow on the ground. The results are presented in column 1 of Table A5. The estimated coefficient of temperature \times distance is 0.000 and not statistically significant, suggesting no protective effect of proximity. That is, as a student moves closer to the university there is no reduction in the sensitivity of their performance to outdoor temperature. Reassuringly, the coefficient on the primary temperature regressor is not meaningfully disturbed.

An issue about the exercise just described is that we observe two distinct addresses for each student. First, an enrolment address used during a student's application to the university. This is almost always the parental or home address. Second, the term address that students are encouraged to keep updated. For some, the application address will be where they actually live, for some it will not, and the lack of variation reflects a failure to update personal details rather than a lack of relocation.

Ideally, we would like a sample of students for which we know where they live with some additional assurance. We construct something close to this in two ways. First, we identify those students who have a term address *distinct* from that at enrolment. We call these students 'movers'.¹⁸ Second, we identify those students who are non-movers but for whom the application address is within 10 km of the university campus. These students live within ready commuting distance of the university and in most cases live at home during their studies, something that is common amongst Canadian undergraduates.

Column 2 reports the results from movers and column 3 from non-movers with an enrolment address within 10 km of campus. The main temperature coefficient of interest remains similar across the three samples, and in each case is statistically significant, despite much eroded sample sizes in column 2 and 3. The coefficients on the temperature \times distance interaction are small and insignificant at conventional levels, discouraging the view that proximity alone delivers a meaningful protective benefit.

¹⁸While it is possible that some families might move in the period between receiving offer and the start of studies, this number is likely small.

In Table A4 we present results of a different approach. We stratify by distance the sample of students who report a term time address within 20 km of campus, irrespective of whether or not they are in our movers sample. In most cases the address that we use is likely the student's residential address. The estimated coefficient on temperature is stable across columns, even in column which estimates only on students who are 'currently' living within 2 km of campus.

Subject to the caveats already noted, the exercises presented in Tables A5 and A4 provide no indication that living close to place of work mitigates the effect of outdoor cold on performance. To the extent that distance correlates with direct exposure to outdoor temperature this implies that it is not the 'amount' of direct exposure which drives the decrement in performance. A similar impact of cold weather is seen even among those who live close to campus. This is more consistent with psychological rather than physiological mechanisms, or other channels identified that do not depend primarily on exposure length.

Apart from location choice, there may be pecuniary ways in which individuals may mitigate the effects of weather to their person. For example, a student may invest in better quality winter clothing, or avoid waiting for a bus by using taxis on particularly cold days. Here we explore a possible role of affluence in temperature-protection.

We do not directly observe the financial circumstances of our sample. However we do know the address reported at first enrolment, which is likely the parental or home address. As a proxy for financial circumstances, we use the average income level at the associated six digit postal code at enrolment as measured in the 2016 Canadian Census. We add this to our preferred specification as an interaction term only, as the student fixed effect will already have accounted for individual income. We present these results in Table A3.

In column 1 we work with all students, including foreign students, provided they had an eligible six digit postal code at enrolment. Because there exists the possibility that the Canadian address reported for a foreign student may a poor indicator of familial wealth, we restrict our sample to domestic students in column 2. In either specification, the main coefficient remains positive and significant. It is somewhat larger than in Table 2, and is now interpreted as the effect on a student from an enrolment address in a hypothetical postal code with average household income of zero dollars. The negative and significant coefficient on the temperature \times average income interaction indicates a protective effect of family affluence. Each 10,000 CAD increase in average

household income in postal code of origin is associated with a 3.7% reduction in the sensitivity of a particular student to cold. A histogram of household incomes is presented in Figure A5. Compared to a zero income benchmark, a student coming from a postal code in the modal category (namely 40,000 to 50,000) benefits from a roughly 15 - 19 % mitigation of cold sensitivity.¹⁹

Overall these exercises are consistent with a protective, but still less than complete, effect of family affluence.

1.7.3 Biological

In this section we present evidence consistent with the results of small scale studies of physiological or psychological adaptation to extreme temperatures mentioned in Section 2. We do this by looking in more detail at the cold-sensitivity of students from other countries and how they evolve over time.

In Table 4 we established that foreign students were statistically more cold-sensitive than domestic students. That Canada is a cold country implies that most students from abroad are from warmer climates. Despite our data not including country of origin at the student level for privacy purposes, we construct a subsample of students most likely to be ‘hot’ countries by leveraging their language of instruction. The University of Ottawa is the largest bilingual English-French university in the world and many undergraduate programs can be taken in their entirety in both languages. As part of its cultural mission the university encourages applications by students from countries of the Francophonie through substantial fee reductions, scholarship programs and promotional efforts.²⁰ 41% of foreign students use French as their language of correspondence with the university. Without knowing individual-level country of origin, the overwhelming majority of non-domestic come from the nations of French Africa (Cote d’Ivoire, Senegal, Cameroon, etc.), or the French Caribbean (Haiti, Dominican Republic etc.) at the aggregate level. These are all hot countries with winter low temperatures typically 25 to 40 degrees Celsius warmer than Ottawa. We identify these students in two ways. First, we

¹⁹Caution should be used in interpreting these results, as the astute reader would note a linear model predicts an income of 267,838 CAD would perfectly offset, and above that reverse, the effects of cold. While we do not see such wealth in our data due to measurement at the postal code (rather than individual) level, it is reasonable to assume that there are diminishing returns to wealth.

²⁰For example foreign students from French-speaking institutions pay domestic rather than foreign fees, which for 2014 - 15 implies a reduction from 22 600 CAD per year to 6,800 CAD.

construct a sample comprising foreign students that elect to study entirely in French across all four years of their program ('Method 1'). Second, reflecting that many students who arrive as unilingual French will develop their English-language skills sufficiently to take at least part of their later studies in English, we relax the sample criterion to comprise foreign students that elect to study only French-taught courses in their first year ('Method 2').

Column 1 in Table 9 reports the result of estimating our preferred specification on the Method 1 subsample, with column 2 estimated on remaining foreign students (most of which come from China and the United States). We can see that the effect of cold on hot country students is much larger than even the effect on international students in general (column 2). The central estimate suggests that a 10°C reduction in outdoor temperature causes a decrement in performance of almost half (45.9%) of a standard deviation. The results in columns 3 and 4 are those estimated on the subsample constructed on the basis of Method 2. They are consistent, though the implied decrement in performance for a 10°C reduction in outdoor temperature is somewhat smaller at 29.9% of a standard deviation.

The results presented to this point have been based on within-student variation in performance under different temperature treatments across their entire period of study. Here we explore how the performance of arrivees changes over time.²¹

The results in Table 10 are estimated only on exams taken during the first year of enrollment. Because this specification incorporates a temperature \times foreign interaction term, the estimated coefficient on temperature, 1.124** represents the effect of temperature on a domestic students, within a course level, during their first exam season. That the coefficient on the temperature \times foreign interaction regressor is positive and significant confirms the earlier finding that foreign students are much more cold-sensitive in their first year.

This exercise is important for another reason. If cold winter temperatures directly affect student attrition rates, then in all specifications we are estimating on temperature 'survivors'. Our results could then be attenuated, particularly at upper course levels. By estimating column 1, we better approximate the effect of cold on performance absent students self-selecting out during the course.

Column 1 is estimated on all students, irrespective of whether they grad-

²¹All specifications include a course-level fixed effect (e.g. second-year or 2000-level courses), to disentangle the effect of course difficulty from the number of years enrolled. The correlation between course level and years enrolled is 0.65.

uate. In column 2 we conduct the same exercise, looking at courses taken in first year of enrollment, but now *only by those students that ultimately graduate*. This is more akin to a balanced panel estimate than the earlier results, and addresses any concern that the propensity to select out of sample during the course of a program might be different between domestic and foreign students. The results here suggest that among domestic students there is indeed disproportionate attrition of cold-sensitive students, as we would expect, but little evidence that the same applies to their foreign counterparts.

To explore adaptation over time, in column 3 we look at all exams taken, but include an interaction term between temperature and number of years enrolled. The exercise is repeated in column 4 where we restrict attention to that subset of students who ultimately graduate. The temperature \times years enrolled coefficients are small and statistically insignificant, indicating that as domestic students spend more time at the university their sensitivity to cold does not change. The large and statistically significant coefficient on the triple interaction term – how foreign student’s sensitivity changes over time – indicates as these students spend more time in Ottawa they become substantially less sensitive to cold. Among both the entire sample and the students who ultimately graduate, the differential between domestic and foreign students is eroded such that it is nearly eliminated after roughly 3 years from their first exam season. This is consistent with the notion of habituation or psychological cold tolerance “... depending largely on the individual’s familiarity with cold” [Enander, 1984].

1.8 Robustness

In Table 11 we challenge the robustness of our main results by re-estimating our preferred specification using alternative temperature measures (corresponding to column 6 in Table 2, which is reproduced in column 1 here).

Alternative temperature metrics The treatment variable of interest throughout the study has been same-day mean temperature. This is calculated as the average of the daily maximum temperature and the minimum temperature. In columns 2 through 5 we replace this measure with alternatives. In column 2 the 24 hour (equally-weighted) daily average temperature, in column 3 the daily minimum temperature, in column 4 exam time temperature, and in column 5 temperature measured at the next closest weather station (Ottawa International Airport, 14 km from the centre of campus). In each case, the qualitative result sustains - cold outdoor temperature causes

a decrement in indoor performance. For comparability between the columns we have also included the mean and standard deviation of the temperature measure applied in each.

Outliers To explore the possibility that the estimated effects are driven by a small number of outliers, we winsorize the treatment variable in column 6. Specifically, we assign the coldest 10% of observations the 10th percentile temperature value and the 10% of warmest observations the 90th percentile value. The results of this exercise are largely the same as our preferred, discouraging the view that our effect is driven by a small number of extreme observations.

Precipitation Throughout the analysis we have been careful to control for the role that precipitation might play, both in its own right and in interaction with temperature. As an additional exercise we reestimate our main specification on the 288,717 exams taken on those days when there was no precipitation ('dry days'). The results are reported in column 7 of Table 11. The sign and significance of the coefficient estimate are sustained, while the coefficient is somewhat larger in value. That we observed the effect even on days absent precipitation provides reassurance that our main specification does a good job of isolating temperature effects from the possible confounding effects of precipitation.

'No controls' specification All of our specifications have included basic controls, for example same-day precipitation. For transparency we report a skeletal specification in which the only regressor is temperature in column 8. Our results sustain.

Placebo As a further test for flaws in our study design that could generate spurious associations between our temperature and performance measures we report here the results of a placebo exercise.

For each student there is vector of exam dates and a vector of associated exam temperatures. To generate placebo temperatures, we separate the two vectors, randomize the order of the exam temperature vector and reattach them. This reassigns temperature treatments randomly without replacement, within-student. Once reattached, recognizing the likely serial correlation within a particular December, we drop any exams for which the randomization assigned a placebo temperature from the same exam period (this necessarily drops any student who writes exams only in a single exam period). The preferred specification is re-estimated with these falsely-assigned treatment values, generating a single coefficient value and associated t-statistic. We repeat this 1,000 times, generating 1,000 temperature coefficient values and 1,000 t statistics. The distributions of these are plotted in Figure 6. It can be seen

that the values derived from the main analysis for both coefficient (0.809) and t statistic (10.408) lie far to the right of any of the placebo-generated values.

Alternative standard errors In Table 12 we report the results of using alternative standard errors for our main analysis. Our main analysis reported standard errors clustered at the student level, corresponding with the panel setting of our data. It is likely that observations within student are correlated (even after accounting for individual fixed effects). Because of this we also apply Huber-White heteroskedasticity robust standard errors. In the second column we provide standard errors that are unclustered and find no meaningful changes in their size. In the third column, we cluster by student cohort, clustering at what could be considered treatment level (for example cold in first year could be different than cold in second year, and cohort determines this inter-year pattern). The challenge here is the low number of cohorts available, forcing us to bootstrap. While the standard errors as measured in this manner are around three times larger, our effect size is still significant at a level well beyond 1%. In column 4 we define treatment levels by exam temperature ventiles and cluster at that level, again with no impact on our conclusions.

1.9 Conclusions

It is obvious that extreme weather can make those working outdoors less productive. However, any link from outdoor temperature to the quantity and quality of work done in indoor, climate-protected environments is potentially crucial in understanding the climate-economy connection, especially in sectors that are not obviously climate-sensitive, such as agriculture.

While a small number of studies have cast light on this question in the case of extreme heat (generally temperatures over about 30°C) we look to the other end of the temperature distribution, finding substantial and apparently robust effects of low outdoor temperature on internal cognitive performance in our setting. That (a) the effect persists even though the students are protected by close-to-perfect climate control, (b) the effect size appears insensitive to the “amount” of exposure that an individual student experiences directly and, (c) sensitivity amongst those new to such temperatures diminishes with repeated exposure, all fit with existing evidence from psychology and biology that the main mechanism or mechanisms at play may be psychological rather than physiological in nature. Our results are consistent with psychological habituation as adaptation, which although less than complete, is able to nullify

the difference in sensitivity between locals and those arriving from warmer climates in the space of around three annual cycles.

The analysis points to a previously unaccounted for benefit of climate change in historically cold places projected in future to experience less cold days. At the same time an unaccounted for cost of climate change in places projected to experience more cold days - in particular those impacted by the weakening of the polar vortex. Additional distribution effects come from secondary results, for example we that men are more sensitive to cold temperatures than women. And the affluent are better insulated from the cold.

Our setting provided the opportunity to conduct a detailed analysis of the scope for adaptation at various loci. While in most cases we found evidence consistent with the protective benefits of adaptation, in no case was the protection complete.

While the performance of university students taking exams is an important social outcome in its own right, the quantitative impacts of the insights of the effect identified depend upon the extent of external validity. If similar decrements in performance were to occur in the workplace, especially in those settings involving high-value, mentally taxing work, the implied economic burden of cold days (alternatively, the *benefits* associated with any reduction in the frequency of cold days) would be large. Investigating the generality of any effects identified here could be a fruitful area of future research.

1.10 Tables and Figures

Table 1: Summary Statistics

	All	Female	Male	Domestic	Foreign
Course Grade	71.98 (10.31)	72.87 (9.82)	70.62 (11.02)	72.27 (10.16)	67.93 (12.22)
Temperature ($^{\circ}\text{C}$)	-5.13 (5.68)	-5.21 (5.68)	-5.01 (5.67)	-5.22 (5.69)	-3.96 (5.59)
Precipitation (mm)	2.12 (4.12)	2.13 (4.14)	2.1 (4.09)	2.13 (4.14)	1.99 (3.77)
Snow on Ground (cm)	2.46 (2.74)	2.46 (2.73)	2.46 (2.75)	2.48 (2.76)	2.17 (2.47)
Foreign	7.43	5.79	9.89	-	100.00
Female	60.00	100.00	-	61.06	46.77
Exams	638,238	384,716	253,522	595,794	42,444
Students	66,715	40,140	26,575	61,814	4,901

Notes: Within-student standard deviations presented. Foreign and female statistics refer to the proportion of exams written by foreign and female students, respectively. Foreign students are classified by immigration status or payment of international student fees.

Table 2: Temperature and Performance (Linear)

	(1)	(2)	(3)	(4)	(5)	(6)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score Preferred
Temperature (°C)	0.833*** (0.043)	0.789*** (0.043)	0.699*** (0.045)	0.750*** (0.047)	0.742*** (0.047)	0.809*** (0.078)
Precipitation	Y	Y	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y	Y	Y
Day of Week FE		Y	Y	Y	Y	Y
Date in Month			Y	Y	Y	Y
Relative Humidity				Y	Y	Y
Snow on Ground					Y	Y
Windchill						Y
Exams	638238	638238	638238	638238	638238	638238
Students	66715	66715	66715	66715	66715	66715

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 3: Temperature and Performance (Non-linear)

	(1) Z-Score	(2) Z-Score	(3) Z-Score	(4) Z-Score	(5) Z-Score	(6) Z-Score Preferred
-15°C	7.670*** (1.091)	4.621*** (1.119)	4.376*** (1.120)	4.091*** (1.123)	4.300*** (1.124)	3.782*** (1.161)
-12.5°C	-2.887** (1.354)	-6.465*** (1.385)	-6.760*** (1.387)	-6.729*** (1.388)	-6.651*** (1.387)	-7.374*** (1.452)
-10°C	14.569*** (1.108)	10.651*** (1.126)	9.296*** (1.170)	8.853*** (1.175)	7.650*** (1.179)	6.756*** (1.296)
-7.5°C	11.446*** (1.140)	7.539*** (1.189)	6.755*** (1.205)	6.377*** (1.210)	3.870*** (1.227)	2.479* (1.497)
-5°C	15.080*** (1.160)	11.653*** (1.167)	11.101*** (1.175)	11.522*** (1.185)	10.807*** (1.186)	9.362*** (1.472)
-2.5°C	16.067*** (1.080)	13.886*** (1.088)	13.020*** (1.105)	13.511*** (1.118)	13.358*** (1.118)	11.565*** (1.557)
0°C	16.788*** (1.214)	14.983*** (1.248)	13.423*** (1.297)	13.875*** (1.309)	14.170*** (1.308)	12.213*** (1.771)
2.5°C	35.063*** (1.637)	33.134*** (1.665)	30.633*** (1.760)	30.903*** (1.763)	32.970*** (1.772)	30.877*** (2.184)
Precipitation	Y	Y	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y	Y	Y
Day of Week FE		Y	Y	Y	Y	Y
Date in Month			Y	Y	Y	Y
Relative Humidity				Y	Y	Y
Snow on Ground					Y	Y
Windchill						Y
Exams	638238	638238	638238	638238	638238	638238
Students	66715	66715	66715	66715	66715	66715

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variables are exam day average temperature bins 2.5 degrees Celsius wide. The reference bin is exam days with temperatures below -15°C. Each bin is separately interacted with precipitation. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 4: Heterogeneity

	(1)	(2)	(3)
	Sex	80 Admission Average	Foreign
Temperature °C	0.927*** (0.091)	0.940*** (0.084)	0.778*** (0.078)
Female=1 × Temperature °C	-0.192*** (0.073)		
80 Admission Average=1 × Temperature °C		-0.311*** (0.072)	
Foreign=1 × Temperature °C			0.486*** (0.162)
Precipitation	Y	Y	Y
Temp × Precip	Y	Y	Y
Day of Week FE	Y	Y	Y
Date in Month	Y	Y	Y
Relative Humidity	Y	Y	Y
Snow on Ground	Y	Y	Y
Windchill	Y	Y	Y
Exams	638238	638238	638238
Students	66715	66715	66715

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. The second independent variable of interest is the interaction between exam day temperature and a subsample identifier. High admission students have an 'A' admission average. Foreign students are classified by immigration status or international fees. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 5: Semester Temperature and Performance

	(1)	(2)	(3)	(4)
	Z-Score	Z-Score	Z-Score	Z-Score
Temperature °C	0.809*** (0.078)	0.798*** (0.078)	0.787*** (0.078)	0.791*** (0.078)
Total HDD Last 30 Days		0.034*** (0.011)		
Total HDD Last 60 Days			0.076*** (0.009)	
Total HDD Last 90 Days				0.057*** (0.012)
Precipitation	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y
Day of Week FE	Y	Y	Y	Y
Date in Month	Y	Y	Y	Y
Relative Humidity	Y	Y	Y	Y
Snow on Ground	Y	Y	Y	Y
Windchill	Y	Y	Y	Y
Exams	638238	638238	638238	638238
Students	66715	66715	66715	66715

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day heating degree days - the number of degrees below 18°C. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 6: Climate Control

	(1) New Building Interaction 1	(2) New Building Interaction 2	(3) Exclude Leaky Rooms
Temperature (°C)	0.837*** (0.081)	0.795*** (0.081)	0.628*** (0.086)
New Building=1	-5.825*** (0.507)	-5.661*** (0.507)	-3.517*** (0.527)
New Building=1 × Temperature (°C)	-0.113* (0.061)	-0.136** (0.061)	-0.031 (0.063)
Course Level FE		Y	Y
Precipitation	Y	Y	Y
Temp × Precip	Y	Y	Y
Year FE	Y	Y	Y
Day of Week FE	Y	Y	Y
Date in Month	Y	Y	Y
Relative Humidity	Y	Y	Y
Snow on Ground	Y	Y	Y
Windchill	Y	Y	Y
Exams	638238	638238	587030
Students	66715	66715	66615

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. The secondary variable of interest is the interaction between a new building (completed after or during 2000 C.E.). Leaky rooms have internal temperature readings outside a $\pm 1^\circ\text{C}$ tolerance band. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 7: Travel to Work

	(1) All	(2) Movers	(3) $\leq 10\text{km}$ Non-Movers
Temperature ($^{\circ}\text{C}$)	0.719*** (0.081)	0.782*** (0.236)	0.866*** (0.288)
Distance (km)	-0.003 (0.004)	-0.002 (0.004)	
Temperature ($^{\circ}\text{C}$) \times Distance (km)	0.000 (0.000)	0.000 (0.000)	-0.025 (0.036)
Precipitation (mm)	-0.437*** (0.060)	-0.438*** (0.170)	-0.980*** (0.297)
Temperature ($^{\circ}\text{C}$) \times Precipitation (mm)	-0.103*** (0.010)	-0.095*** (0.032)	-0.121*** (0.025)
Distance (km) \times Precipitation (mm)	-0.000 (0.000)	-0.000 (0.000)	0.084* (0.044)
Snow on Ground (cm)	-0.513*** (0.054)	-0.631*** (0.161)	-0.161 (0.232)
Distance (km) \times Snow on Ground (cm)	0.000 (0.000)	-0.000 (0.000)	-0.043 (0.032)
Day of Week FE	Y	Y	Y
Date in Month	Y	Y	Y
Relative Humidity	Y	Y	Y
Windchill	Y	Y	Y
Exams	598407	81347	107380
Students	62596	8530	11514

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. The second independent variable of interest is the interaction between temperature and distance to student address (measured in km). Movers are students whose term address is different than their enrolment address. Students whose addresses are never more than 50km from campus. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 8: Family Affluence Proxy

	(1) All	(2) Domestic
Temperature ($^{\circ}\text{C}$)	0.991*** (0.123)	0.939*** (0.127)
Temperature ($^{\circ}\text{C}$) \times Avg. Income	-0.037** (0.018)	-0.038** (0.019)
Precipitation	Y	Y
Temp \times Precip	Y	Y
Day of Week FE	Y	Y
Date in Month	Y	Y
Relative Humidity	Y	Y
Snow on Ground	Y	Y
Windchill	Y	Y
Exams	627352	588005
Students	65404	60962

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. The second independent variable of interest is the interaction between temperature and average income of student address at enrolment (from 2016 Census data). Average income measured in 10,000's CAD. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 9: Heterogeneity Among Arrivees

	(1) Method 1 Probably Hot	(2) Method 1 Other	(3) Method 2 Probably Hot	(4) Method 2 Other
Temperature (°C)	4.591*** (1.048)	1.226*** (0.401)	2.992*** (0.796)	1.495*** (0.425)
Precipitation	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y
Day of Week FE	Y	Y	Y	Y
Date in Month	Y	Y	Y	Y
Relative Humidity	Y	Y	Y	Y
Snow on Ground	Y	Y	Y	Y
Windchill	Y	Y	Y	Y
Exams	6308	36136	9907	32537
Students	985	3916	1275	3626

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. The first column estimates the preferred specification on international students who take all of their courses in French. The second column estimates our preferred specification on international students who took none (N=3,085), or some fraction of their studies (N=981) in French. In the third and fourth column we relax our definition of probably hot country students to those who take all of their first year courses in French. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 10: Adaptation of Arrivees Over Time

	(1) All (Year 1 Exams)	(2) Graduates (Year 1 Exams)	(3) All	(4) Graduates
Temperature (°C)	1.124*** (0.129)	0.498*** (0.163)	0.695*** (0.087)	0.358*** (0.098)
Foreign=1 × Temperature (°C)	0.855*** (0.318)	1.300*** (0.452)	0.866*** (0.252)	1.100*** (0.326)
Temperature (°C) × Years Enrolled			0.022 (0.031)	0.035 (0.033)
Foreign=1 × Temperature (°C) × Years Enrolled			-0.237* (0.139)	-0.394** (0.159)
Course Level FE	Y	Y	Y	Y
Precipitation	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y
Day of Week FE	Y	Y	Y	Y
Date in Month	Y	Y	Y	Y
Relative Humidity	Y	Y	Y	Y
Snow on Ground	Y	Y	Y	Y
Windchill	Y	Y	Y	Y
Exams	265804	136319	638238	426583
Students	66447	33228	66715	33322

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. All specifications include course-level fixed effects (e.g. 2000 level courses). Years enrolled begins at 0 for the first winter of exams, and typically ends at 3 years. In columns 1 and 2, we estimate only on the first year's course results and do not include year fixed effects. Columns 3 and 4 include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 11: Robustness

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean Temp (Preferred)	24 Hour Average	Min Temp	Exam Temp	Next Station	Winsorized at 10%	Dry Days	No Controls
Temp. Measure	0.809*** (0.078)	0.771*** (0.080)	0.386*** (0.064)	0.603*** (0.095)	0.532*** (0.075)	0.724*** (0.094)	0.963*** (0.128)	1.526*** (0.035)
Precipitation	Y	Y	Y	Y	Y	Y		
Temp \times Precip	Y	Y	Y	Y	Y	Y		
Day of Week FE	Y	Y	Y	Y	Y	Y	Y	
Date in Month	Y	Y	Y	Y	Y	Y	Y	
Relative Humidity	Y	Y	Y	Y	Y	Y	Y	
Snow on Ground	Y	Y	Y	Y	Y	Y	Y	
Windchill	Y	Y	Y	Y	Y	Y	Y	
Mean of Measure	-5.14	-4.76	-8.68	-4.16	-5.3	-5.12	-6.64	-5.14
SD of Measure	6.61	6.45	7.51	6.52	6.75	5.52	6.66	6.61
Exams	638238	638238	638238	638238	638238	638238	288717	638238
Students	66715	66715	66715	66715	66715	66715	64016	66715

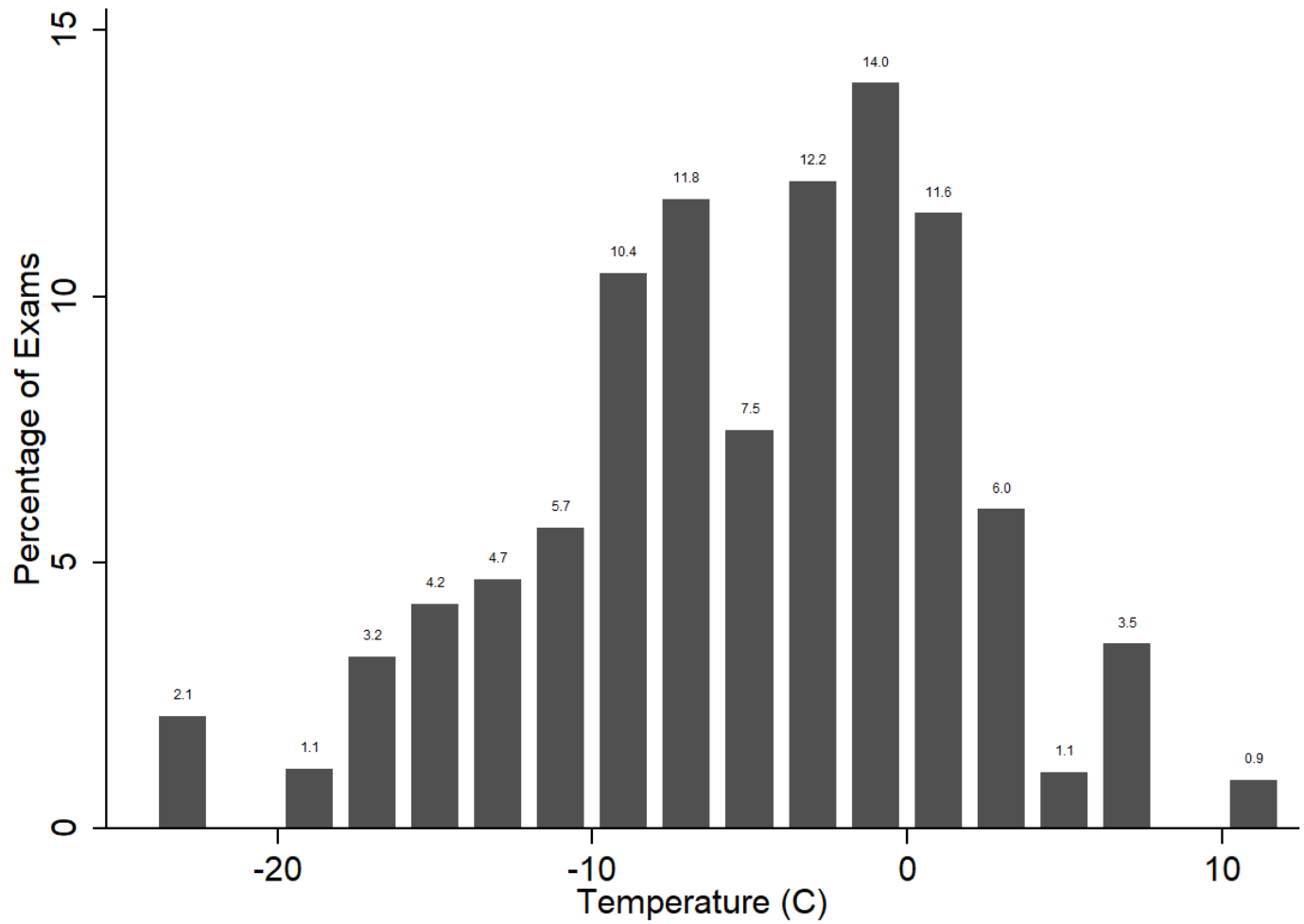
The dependent variable is hundredths of a standard deviation in final exam grade. Each column title denotes the primary independent variable. The first column is average exam day temperature in degrees Celsius, calculated as the average of daily maximum and minimum. The second column is the 24 hour equally weighted average temperature. The third column uses daily minimum temperature. The fourth column uses the average hourly temperature during the 3 hour window of the exam. The fifth column uses daily average temperature from the next-closest weather station (an international airport approximately 14km away). The sixth uses temperatures winsorized at the 10% and 90% level. The seventh column estimates the preferred specification only on days without precipitation. The eighth column simply regresses performance and temperature. Other than ‘no controls’, all specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 12: Alternative Standard Errors

	(1) Preferred Student	(2) Unclustered	(3) Cohort (Bootstrap)	(4) Exam Ventiles
Temperature (°C)	0.809*** (0.078)	0.809*** (0.078)	0.809*** (0.252)	0.809*** (0.225)
Precipitation	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y
Day of Week FE	Y	Y	Y	Y
Date in Month	Y	Y	Y	Y
Relative Humidity	Y	Y	Y	Y
Snow on Ground	Y	Y	Y	Y
Windchill	Y	Y	Y	Y
Exams	638238	638238	638238	638238
Clusters	66715		9	20

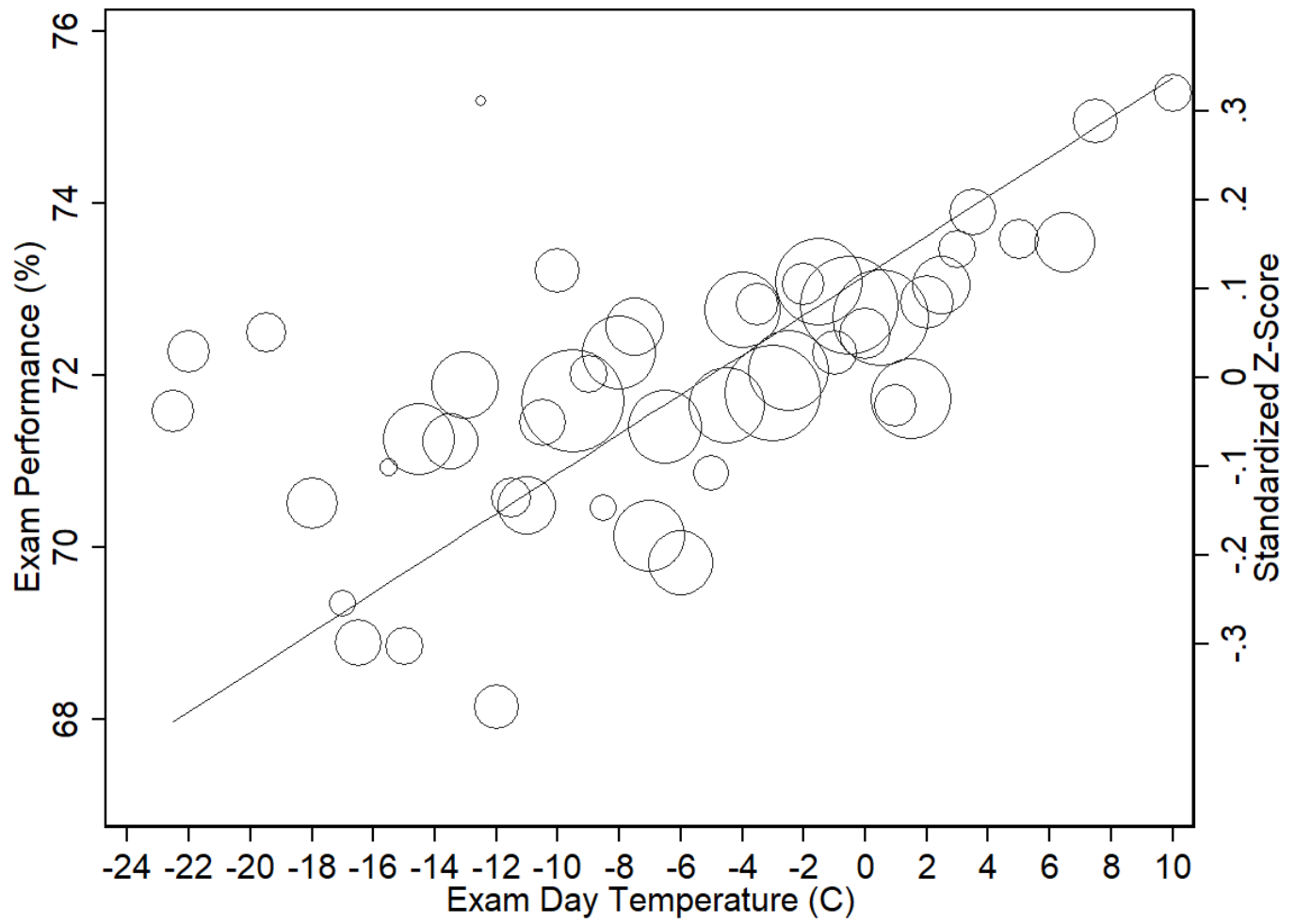
The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. (1) errors clustered at the student level (2) unclustered errors (3) bootstrapped errors clustered by cohort (4) ventiles of average exam temperatures. All specifications include year fixed effects. Within-student fixed effects model. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Figure 1: Distribution of Temperature Treatments



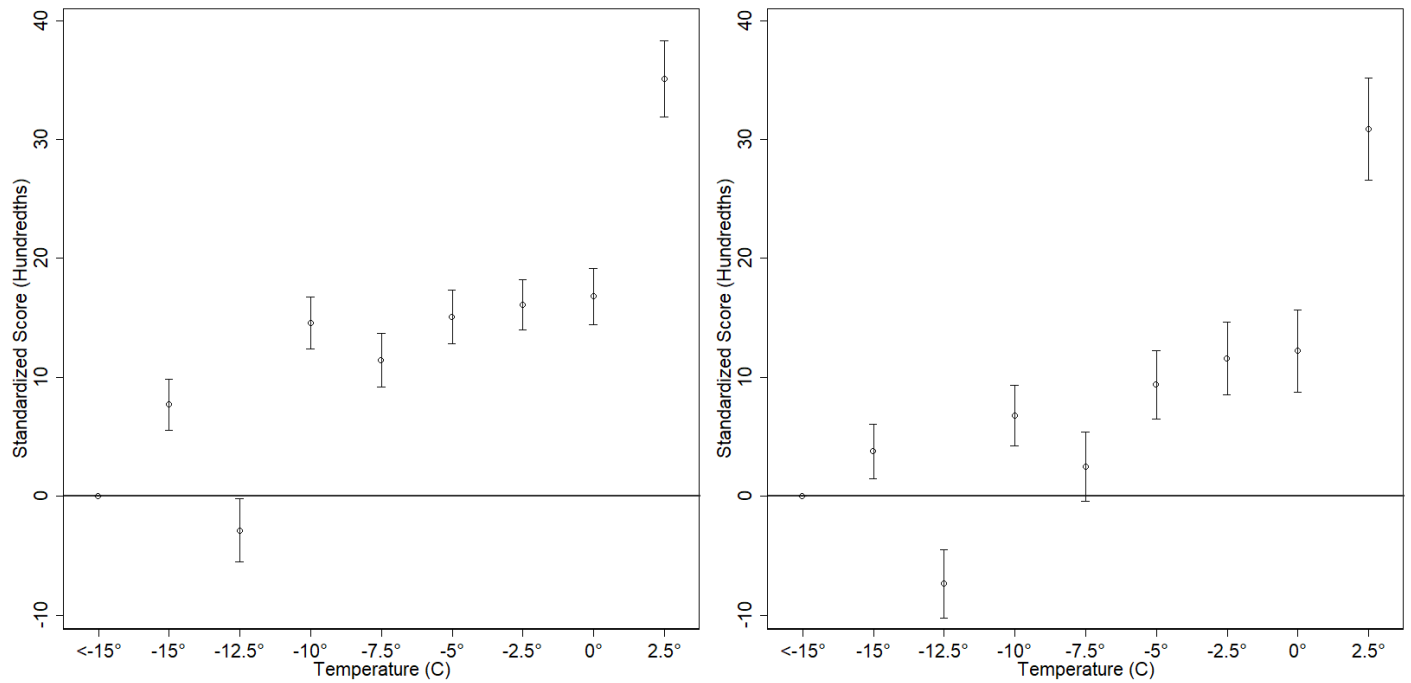
In this figure, we plot the percentage of exams written on days with average temperatures divided into 2 degree Celsius bins. Each exam, rather than each exam day, represents a single observation.

Figure 2: Temperature and Performance (Only Year Controls)



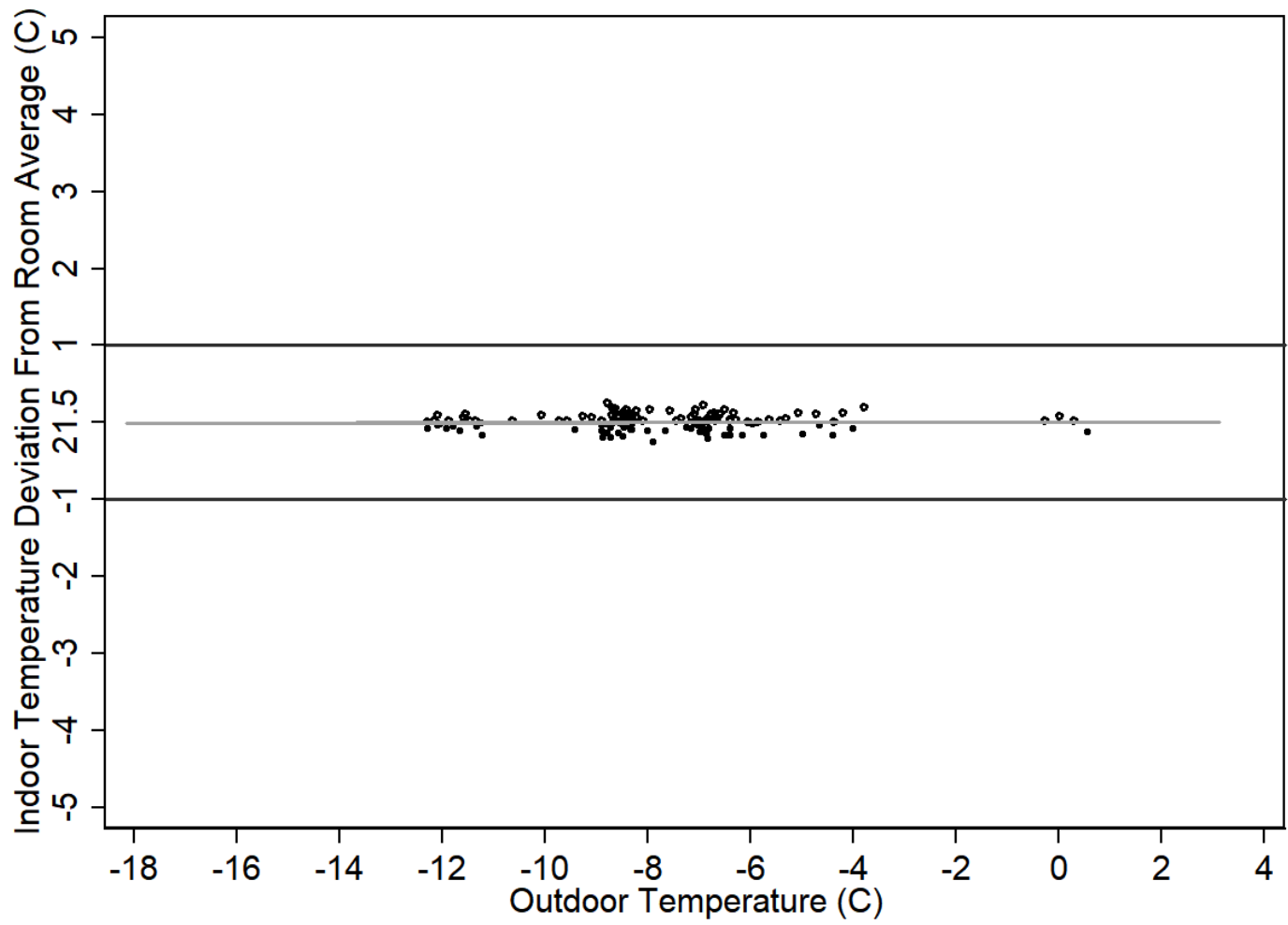
In this figure, we plot the imputed residual exam grade (after accounting for year fixed effects) by exam day temperature. Temperature is rounded to the nearest 0.5°C . Markers are sized proportional to number of observations they represent.

Figure 3: Temperature and Performance (Non-Linear)



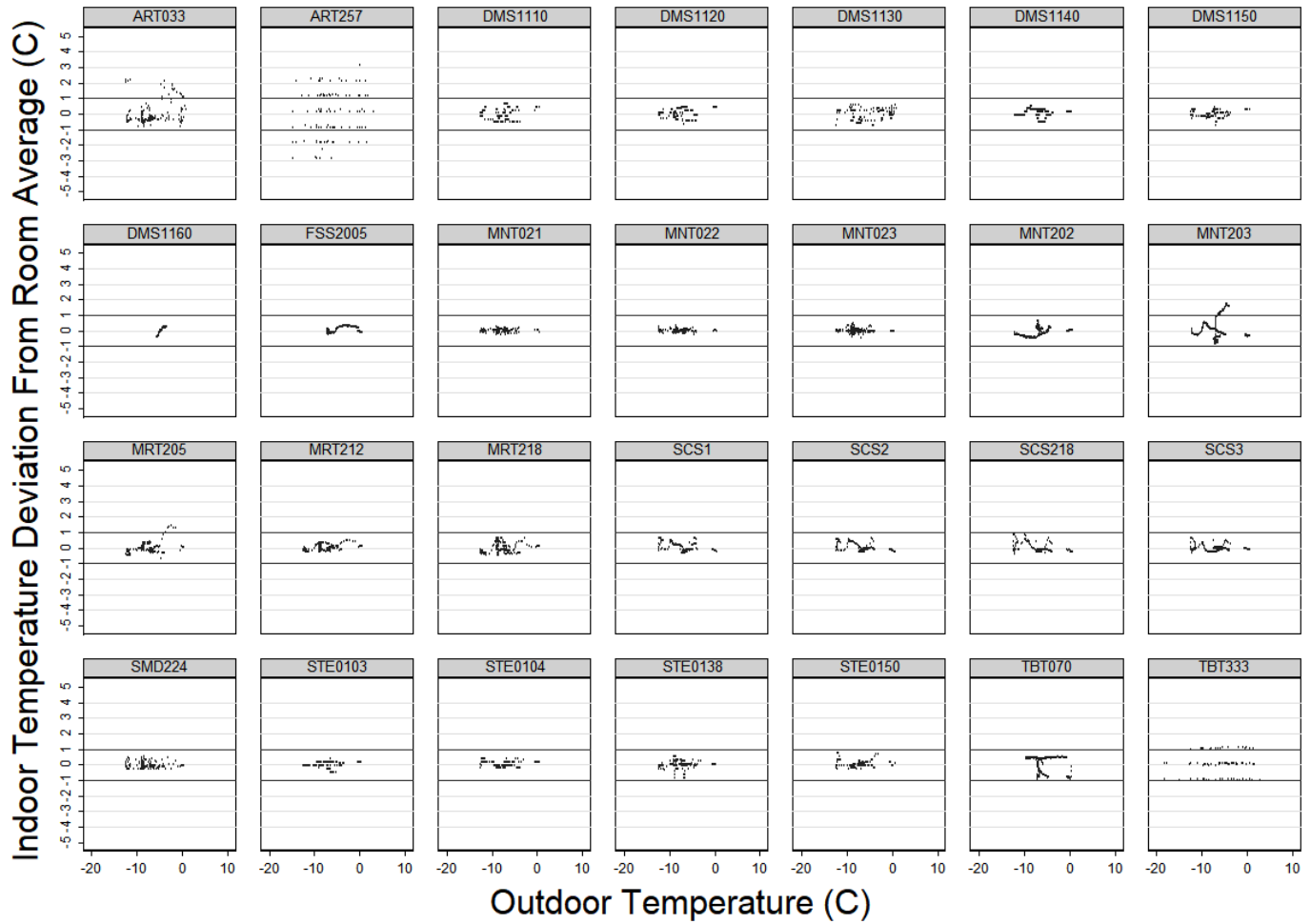
In this figure, we present the estimated coefficients by “binning” daily temperatures into 2.5°C intervals. The reference category is exams written with daily temperatures below -15°C . The dependent variable is exam score standard deviations in hundredths. The left panel corresponds to a parsimonious specification with student and year fixed effects, precipitation, and its interaction with each temperature bin. The right panel corresponds to our preferred specification with additional controls. Whiskers indicate the 95% confidence level.

Figure 4: Indoor and Outdoor Temperatures (MNT021)



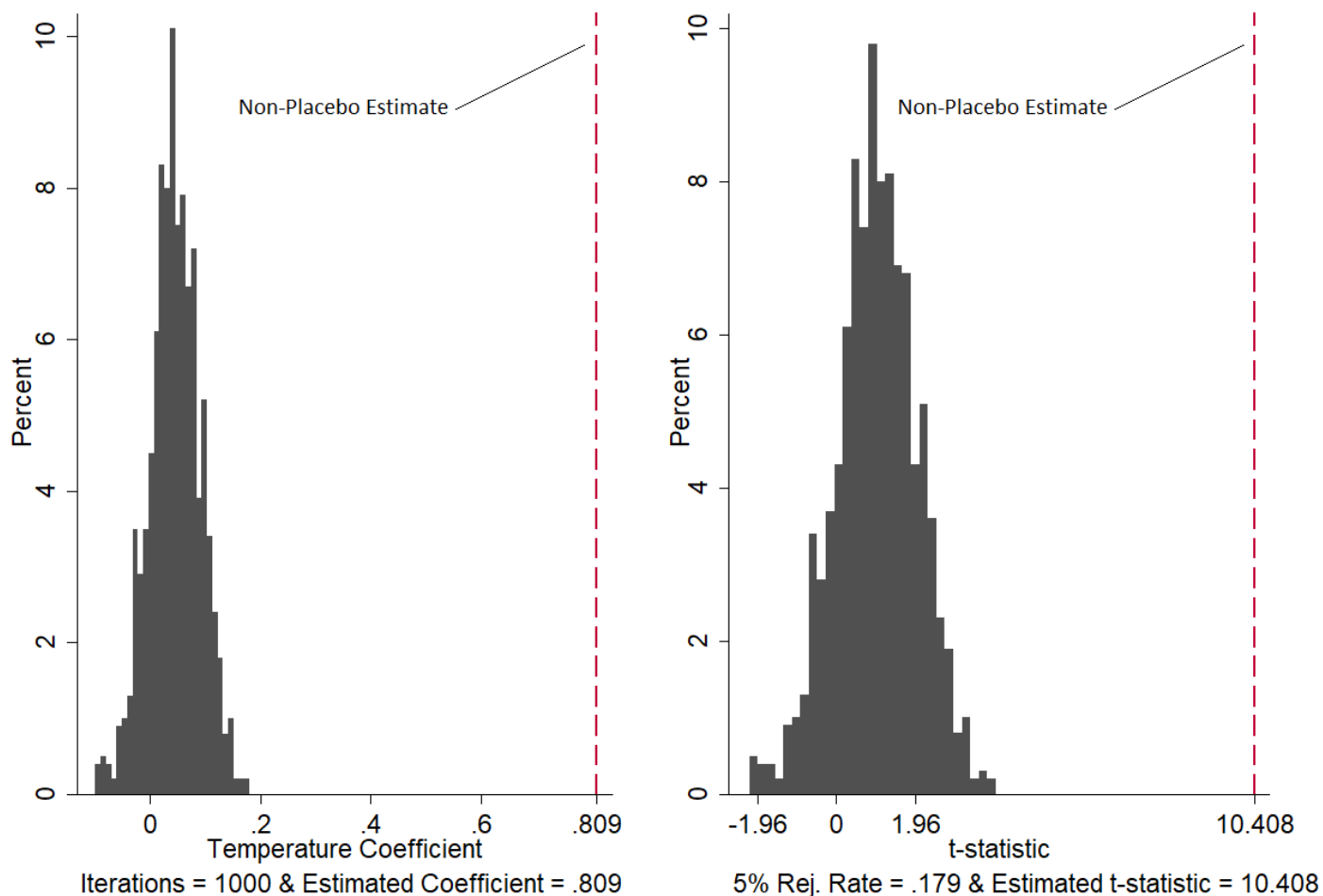
In this figure, we plot the outdoor temperatures realized during 2018 December exams and the internal temperature variations from room average in the largest exam room by contribution to sample. We fit a regression line with slope coefficient of 0.0003 and an associated t-statistic of 0.10. Reference lines are provided at 1°C (above) and -1°C (below) room average.

Figure 5: Indoor and Outdoor Temperatures (Room by Room)



In this figure, we plot the outdoor temperatures realized during 2018 December exams and the internal temperature variations from room average by exam room. Reference lines are provided at 1°C (above) and -1°C (below) room average.

Figure 6: Placebo



In this figure, we present histograms of the estimated temperature coefficients and associated t-statistics for a placebo exam day temperature. Placebo temperatures are randomized within-student and without replacement. If an exam was assigned a placebo temperature from the same exam season, that observation was dropped. The preferred specification in Table 2 was run 1,000 times. A reference line corresponding to our preferred specification, on the correct exam day temperature, is provided in each panel.

1.11 Appendices

Table A1: Temperature and Performance (Linear)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score Preferred
Temperature (°C)	0.609*** (0.040)	0.688*** (0.042)	0.833*** (0.043)	0.789*** (0.043)	0.699*** (0.045)	0.750*** (0.047)	0.742*** (0.047)	0.809*** (0.078)
Precipitation		-0.387*** (0.046)	-0.712*** (0.050)	-0.451*** (0.052)	-0.425*** (0.052)	-0.355*** (0.055)	-0.419*** (0.055)	-0.425*** (0.055)
Temp × Precip			-0.128*** (0.010)	-0.117*** (0.010)	-0.107*** (0.010)	-0.112*** (0.010)	-0.105*** (0.010)	-0.104*** (0.010)
Date in Month					-0.353*** (0.053)	-0.315*** (0.054)	-0.015 (0.062)	-0.013 (0.062)
Relative Humidity						-0.077*** (0.019)	-0.025 (0.019)	-0.019 (0.020)
Snow on Ground							-0.500*** (0.051)	-0.503*** (0.051)
Windchill								-0.037 (0.034)
Day of Week FE				Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Exams	638238	638238	638238	638238	638238	638238	638238	638238
Students	66715	66715	66715	66715	66715	66715	66715	66715

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table A2: Temperature and Performance (Deciles)

	(1) Z-Score	(2) Z-Score	(3) Z-Score	(4) Z-Score	(5) Z-Score	(6) Z-Score Preferred
-14.7°C	6.805*** (1.030)	4.484*** (1.051)	4.277*** (1.052)	3.796*** (1.057)	2.916*** (1.061)	1.733 (1.117)
-10.6°C	8.666*** (1.113)	3.514*** (1.148)	1.960* (1.168)	1.376 (1.171)	-1.192 (1.206)	-3.243** (1.363)
-8.3°C	18.341*** (1.075)	14.276*** (1.102)	12.494*** (1.131)	11.755*** (1.138)	9.297*** (1.170)	6.572*** (1.431)
-6.5°C	8.474*** (1.211)	5.319*** (1.234)	5.282*** (1.234)	5.113*** (1.234)	3.070** (1.256)	-0.023 (1.585)
-4.1°C	19.629*** (1.317)	12.205*** (1.355)	11.748*** (1.358)	12.589*** (1.367)	12.015*** (1.368)	8.621*** (1.715)
-2.7°C	11.557*** (1.127)	7.405*** (1.137)	5.992*** (1.155)	6.711*** (1.163)	5.041*** (1.177)	1.005 (1.702)
-.7°C	17.268*** (1.184)	14.313*** (1.199)	13.270*** (1.208)	14.719*** (1.240)	13.635*** (1.244)	9.348*** (1.802)
.3°C	15.850*** (1.208)	14.862*** (1.245)	12.868*** (1.275)	13.693*** (1.287)	12.730*** (1.291)	8.327*** (1.862)
2.2°C	27.861*** (1.534)	23.528*** (1.574)	20.719*** (1.627)	21.947*** (1.649)	21.914*** (1.649)	17.239*** (2.182)
Precipitation	Y	Y	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y	Y	Y
Day of Week FE		Y	Y	Y	Y	Y
Date in Month			Y	Y	Y	Y
Relative Humidity				Y	Y	Y
Snow on Ground					Y	Y
Windchill						Y
Exams	638238	638238	638238	638238	638238	638238
Students	66715	66715	66715	66715	66715	66715

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variables are exam day average temperature deciles. The reference bin is exam days with temperatures below -14.7°C. Each bin is separately interacted with precipitation. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table A3: Semester Temperature and Performance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score
Temperature (°C)	0.809*** (0.078)	0.796*** (0.091)	0.964*** (0.083)	0.839*** (0.080)	0.798*** (0.078)	0.786*** (0.078)	0.803*** (0.078)
Avg. Temp. Last 1 Days		0.017 (0.058)					
Avg. Temp. Last 3 Days			-0.307*** (0.065)				
Avg. Temp. Last 5 Days				-0.108 (0.080)			
Avg. Temp. Last 30 Days					-1.028*** (0.318)		
Avg. Temp. Last 60 Days						-4.580*** (0.559)	
Avg. Temp. Last 90 Days							-4.395*** (1.152)
Precipitation	Y	Y	Y	Y	Y	Y	Y
Temp × Precip	Y	Y	Y	Y	Y	Y	Y
Day of Week FE	Y	Y	Y	Y	Y	Y	Y
Date in Month	Y	Y	Y	Y	Y	Y	Y
Relative Humidity	Y	Y	Y	Y	Y	Y	Y
Snow on Ground	Y	Y	Y	Y	Y	Y	Y
Windchill	Y	Y	Y	Y	Y	Y	Y
Exams	638238	638238	638238	638238	638238	638238	638238
Students	66715	66715	66715	66715	66715	66715	66715

The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day temperature. The secondary independent variable is average temperature leading up to exam day. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table A4: Travel to Work (Subsamples)

	(1)	(2)	(3)	(4)
	$\leq 2\text{km}$	$\leq 5\text{km}$	$\leq 10\text{km}$	$\leq 20\text{km}$
Temperature ($^{\circ}\text{C}$)	0.919*	0.725**	0.770***	0.834***
	(0.532)	(0.359)	(0.214)	(0.174)
Precipitation	Y	Y	Y	Y
Temp \times Precip	Y	Y	Y	Y
Day of Week FE	Y	Y	Y	Y
Date in Month	Y	Y	Y	Y
Relative Humidity	Y	Y	Y	Y
Snow on Ground	Y	Y	Y	Y
Windchill	Y	Y	Y	Y
Exams	14182	31379	88217	113229
Students	1966	3699	9771	11618

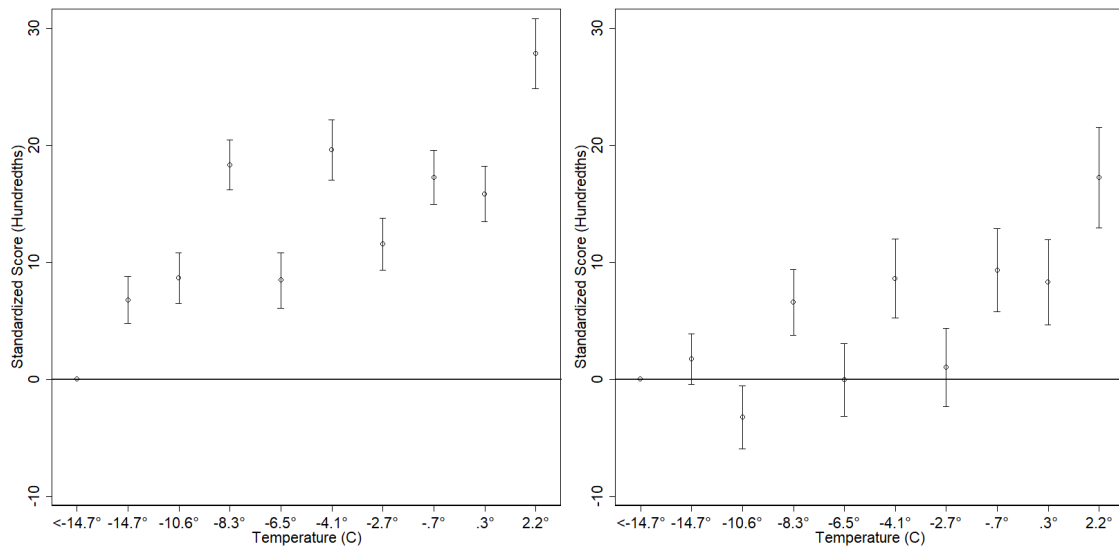
The dependent variable is hundredths of a standard deviation in final exam grade. The primary independent variable is exam day average temperature in degrees Celsius. Each column header indicates the outer radius of successively distant donut-shaped regions. The second column estimates our effect for addresses 2.0 km to 5.0 km from campus. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table A5: Temperature and Performance, Alternative Standardization

	(1) Z-Score	(2) Z-Score	(3) Z-Score	(4) Z-Score	(5) Z-Score	(6) Z-Score Preferred
Temperature (C)	0.730*** (0.043)	0.689*** (0.044)	0.524*** (0.046)	0.605*** (0.047)	0.604*** (0.047)	1.039*** (0.079)
Precipitation	Y	Y	Y	Y	Y	Y
Temp \times Precip	Y	Y	Y	Y	Y	Y
Day of Week FE		Y	Y	Y	Y	Y
Date in Month			Y	Y	Y	Y
Relative Humidity				Y	Y	Y
Snow on Ground					Y	Y
Windchill						Y
Exams	638185	638185	638185	638185	638185	638185
Students	66713	66713	66713	66713	66713	66713

The dependent variable is hundredths of a standard deviation in final exam grade, standardized by year and course. The primary independent variable is exam day average temperature in degrees Celsius. All specifications include year fixed effects. Within-student fixed effects model. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. The sample comprises all exams written in December from 2007-2015. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Figure A1: Temperature and Performance (Deciles)



In this figure, we present the estimated coefficients for indicator variables created by assigning daily temperatures into decile intervals. The reference category is exams written in the 10% coldest daily average temperatures. The dependent variable is exam score standard deviations in hundredths. The left panel corresponds to a parsimonious specification with student and year fixed effects, precipitation, and its interaction with each temperature bin. The right panel corresponds to our preferred specification with additional controls. Whiskers indicate the 95% confidence level.

Figure A2: Precipitation

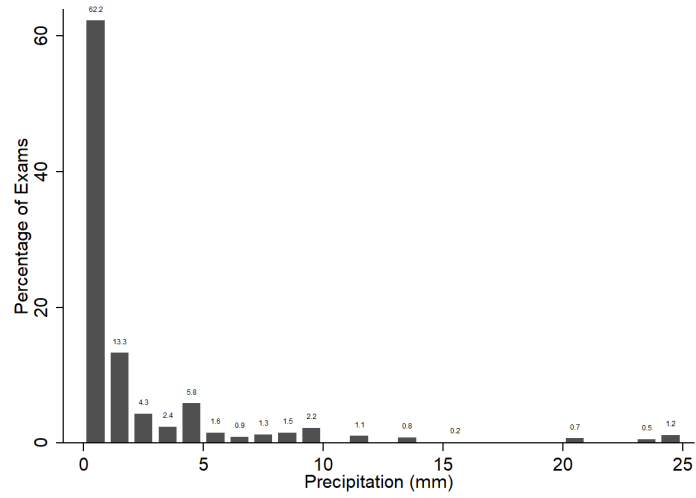


Figure A3: Snow on Ground

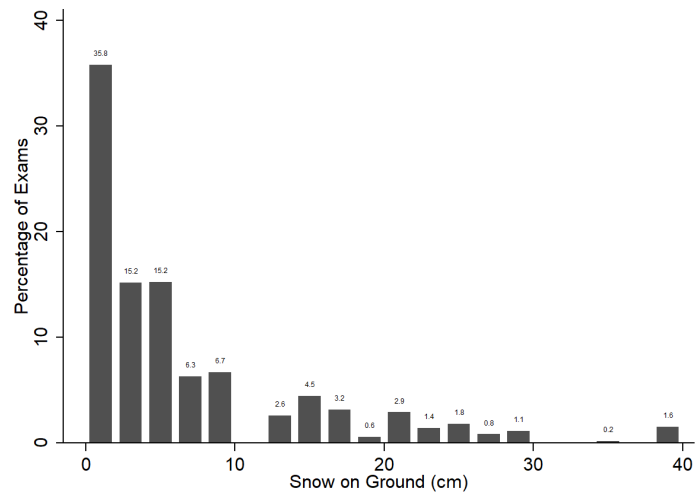


Figure A4: Distance to Student Address

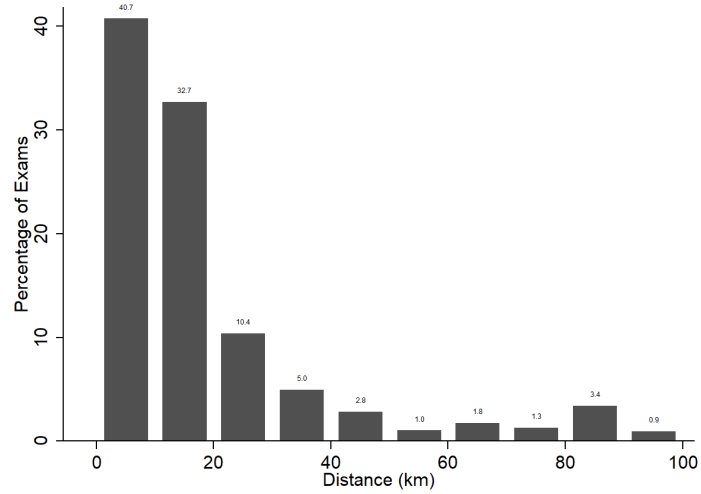
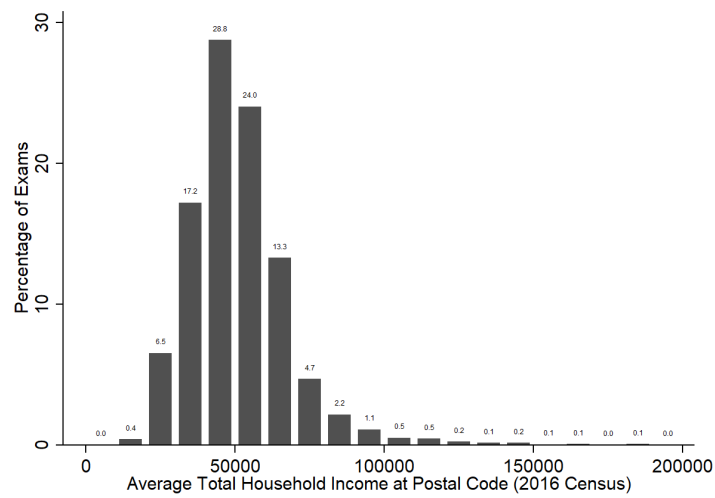


Figure A5: Student Application Address Average Income



Chapter 2

A Boss Like Me: Employees' Preferences for Employers

2.0.1 Abstract

Extensive evidence points to employers discriminating against prospective employees on grounds of race and gender. In a choice experiment conducted in an online labor market setting we analyze 14,544 job offer choices made by 909 job-seekers, finding the first systematic evidence of race bias in labor *supply*. In our preferred specification white respondents are 11.3% more likely to choose a job offered to them by a white manager than an otherwise identical job offered by a black. In contrast black respondents are 17.3% less likely. We apply probabilistic discrete choice modeling methods to estimate intensity of preferences over race of manager at group- and individual-level, and find them substantial. There is little evidence of bias on basis of gender. The results, if sustained in the population at large, point to a further channel through which particular race-gender types could be disadvantaged in a competitive market for managerial roles.

2.0.2 Thanks

We are grateful to Abel Brodeur, John List, Steve Martin, Alberto Salvo and diverse seminar participants for useful advice. Errors are ours.

2.0.3 Ethics and Collaboration

This research was completed under University of Ottawa Research Ethics Board file number 03-17-12. It was completed in collaboration with Professor Anthony Heyes. The student's contributions include but are not limited to conception of research question, creating online experimental instrument, and drafting of chapter.

2.1 Introduction

Does race or gender still matter in the modern American office? A large experimental literature suggests that employers still hire based on employee race and gender. But do *employees* make application decisions based on *employers'* race or gender? This experiment is the first to find employee preferences for employers of own race and gender.

Discrimination on grounds of race and gender is commonly encountered in the labor market and other settings.¹ Popular commentary and a large body of evidence contend that employers treat black and female job applicants less favorably than their white and male counterparts. Altonji and Blank [1999] provide a thorough review of the literature, and produce earnings regressions detailing the black and female wage gaps. For experimental evidence, Bertrand and Mullainathan [2004] found that a fictive CV sent to help-wanted adverts in Chicago and Boston was around 30% less likely to attract a callback for an interview if it carried an African-American sounding name rather than a white-sounding name.² A number of related studies explore gender.³

We conduct a choice experiment in the online labor market Mechanical Turk. Under the guise of a study to explore worker preferences over job characteristics, respondents are asked to choose between pairs of offers for

¹Arrow [1998] enumerates: social relations, residential location, legal barriers, income, wages, prices paid and credit extended. Neumark [2016] provides a comprehensive review of the literature. Bertrand and Duflo [2017], Lane [2016] detail field and lab experiments. Example lab experiments include prisoner's dilemma [Charness et al., 2007], minimum effort game [Chen and Chen, 2011] and norm enforcement [Goette et al., 2012].

²Subsequent experiments provided consistent results. Bertrand et al. [2005] recruited 115 subjects and asked them to construct a hypothetical short-list of 15 out of 50 CVs that were the 'best fit' for a company in filling a position as an administrative assistant. In a related online field experiment in Toronto, Oreopoulos [2011] found similar effects in Toronto when switching from a CV with a common English name to one with an ethnic name.

³We explore both gender and race-based preferences. Our main results relate to race.

office-based jobs that vary in several dimensions (hourly wage, level of independence, opportunities for advancement). We then manipulate the race and gender of the offering manager by varying the name and photograph in the e-mail signature.

Analyzing the 14,544 job offer choices made by 909 respondents, we find that white respondents are up to 11.3 % more likely to choose a job offered to them by a white manager. These findings are reversed for black respondents, who are up to 17.3 % *less* likely to accept a job offer from a white manager. Effects related to gender are much smaller and in most cases insignificant. To measure *intensity* of preferences, we use a suite of standard discrete choice models to estimate group and individual level willingness to pay (WTP) for manager characteristics. In our preferred specification, 23% of white respondents are willing to pay more than 0.50 USD per hour for a white manager. In contrast, 39% of black males and 49% of black females are willing to pay more than 0.50 USD per hour for a black manager.

Previous research into race and the labor market has focused on labor demand - how the characteristics of the job-seeker affect his or her ability to find a job and on what terms. We believe the results presented here are the first evidence of bias by prospective *employees* on the grounds of race (and to a lesser extent gender) of the person making the job offer.⁴ The general tone of our results, that applicants prefer to work for those similar to themselves, is consistent with recent findings of own-group preferences of coworkers [Hedegaard and Tyran, 2018] and evaluators [Feld et al., 2015].

The remainder of this chapter is structured as follows. Section 2 introduces the experimental design. Section 3 describes the estimation strategy. Section 4 contains the main results. Section 5 contains the results of robustness exercises. Section 6 concludes.

⁴A small literature uses choice experiments to uncover applicants' preferences for job benefits such as flexible working hours. Their WTP estimates also vary substantially between and within demographic groups [Eriksson and Kristensen, 2014, Wiswall and Zafar, 2016, Mas and Pallais, 2017]. Job applicants have also been shown to sort themselves by gender according to stated and inferred job offer information ([Flory et al., 2014, Kuhn and Villeval, 2015]). We extend this line of research by considering the role of manager race and gender on the attractiveness of a job offer.

2.2 Methods

We use a choice experiment to elicit respondent preferences over hypothetical job offers. Discrete choice experiments have been widely-used. Examples include public preferences over alternative bundles of environmental goods [Hanley et al., 1998], food safety [Finn and Louviere, 1992] and health interventions [Bansback et al., 2012]. The closest application of a choice experiment to our study is Eriksson and Kristensen [2014], who investigate job applicants' preferences for job amenities such as work hours flexibility. Our results prove robust to various ways of analyzing the data. Our main results are derived from multinomial logit analysis, but sustain under multinomial probit or mixed logit alternatives.⁵

The online experiment was conducted using Mechanical Turk (MT). In each task, a respondent indicated which hypothetical job offer they would prefer from two presented. In this way each respondent 'ranked' 20 job offer pairs. Most offers embodied randomized characteristics such that pairs were idiosyncratic to respondent. However, four of the pairs were 'fixed' and answered by every respondent at the same point in the series. Once finished, respondents completed a demographic survey. Workers were paid 0.25 USD per response and the median time to complete the experiment was approximately 6 minutes.⁶

2.2.1 The Setting: Amazon's Mechanical Turk

Respondents were recruited from Amazon's Mechanical Turk (MT). MT is a large crowd-sourcing internet marketplace that connects individuals and businesses. Requesters post 'Human Intelligence Tasks'. HITs are generally quite small, requiring just a few minutes of time. Workers, commonly referred to as Turkers, browse and complete the posted tasks in exchange for payment.

⁵See [McFadden, 1974, Hausman and McFadden, 1984, McFadden, 1986, McFadden and Train, 2000] A closely related method is conjoint analysis. For some discussion of the comparative properties and merits of the alternatives see Louviere et al. [2010].

⁶Our hourly rate of 2.50 is above the 1.71 USD per hour used previously to replicate lab results [Paolacci et al., 2010]. A number of studies have shown that the quality of responses obtained on MT is insensitive to the rate of remuneration [Amir et al., 2012]. To address the risk of under-motivated respondents failing to give adequate attention to our tasks, we include an attention check following Mas and Pallais [2017]. We also test (and reject) the hypothesis that fatigue or wandering attention leads to answers that vary systematically with position in series.

There are a confirmed 500 000 Turkers worldwide.⁷ With an estimate of 78% in the United States as of February 2018.⁸ MT allows the requester to restrict respondents based on their location or past record of satisfied requesters.

MT is well-suited to the conduct of survey experiments [Kim and Hodgins, 2017, Kees et al., 2017, Horton et al., 2011, Buhrmester et al., 2011, Paolacci et al., 2010] provide evidence of the quality of responses obtained in varied research settings, one concluding that “... the data obtained are at least as reliable as those obtained via traditional methods. Overall, MT can be used to obtain high-quality data inexpensively and rapidly” [Buhrmester et al., 2011].

In terms of sample caution is needed. MT workers in the US over-represent some groups (in particular Hispanic females, young Asian males and females) and under-represent others (African-Americans of all ages), [Huff and Tingley, 2015]. However we will only present results estimated *within* race-gender types. Importantly for us, Huff and Tingley [2015] and others present evidence that *within* socio-demographic groups (defined by race, gender and age) personality measures and responses to questions probing social and political attitudes are similar to those from population-representative survey platforms such as the Cooperative Congressional Election Survey. Clifford et al. [2015] assessed personality and value-based responses from MT-recruited samples and compared them to the same responses derived from two widely-used benchmark US national samples (one online and one face-to-face) and found that “all three samples produced substantively identical results” [Clifford et al., 2015].⁹ While these findings are reassuring, we are nonetheless cautious with regards to external validity. Probing how far our findings extrapolate to the wider population is an important question in future research.

The platform is now commonly used for both experimental and survey completion in a range of disciplines.¹⁰ Keith et al. [2017] provide an excellent meta-study of MT-based research in organizational research, along with

⁷<https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/OverviewofMturk.html>

⁸<http://demographics.mturk-tracker.com/#/gender/all>

⁹A number of researchers have run parallel studies on MT and traditionally-sourced samples (such as national telephone panels and university laboratory pools) and found no significant difference in patterns of response. Examples include Paolacci et al. [2010] using a series of standard judgment and decision making tasks, Simons and Chabris [2012] on scientific beliefs, and Behrend et al. [2011] who assessed the ‘big 5’ personality traits of respondents using the popular 20 point International Personality Item Pool.

¹⁰Feldman(2017) concludes: “MT is a very powerful tool for quick and inexpensive data collection. There are lots of high profile articles popping up in various journals across many domains that have come to the same conclusions as I have - MT is an important tool.”

recommendations for execution, many of which we adopt. Recent applications in economics include (1) Kuziemko et al. [2015] who used randomized survey experiments to elicit preferences in the US over income inequality, redistribution and economic growth; (2) DellaVigna and Pope [2016] explore how various monetary and non-monetary motivators impact worker effort; (3) In their study of recommender systems, Yeomans et al. [2017] ask Turkers to evaluate entries from a database of jokes; (4) List and Momeni [2017] explore preferences of respondents over prospective employers with different (experimentally-manipulated) corporate social responsibility profiles.

2.2.2 Job Offers

Respondents chose between pairs of hypothetical offers of employment as an administrative assistant. The role is one of the most common positions in the US and other developed economies. For each job they were shown a mock ‘e-mail of offer’, with pairs shown side by side in their web-browser. Figure 1 presents a typical pair.

Each job varies along dimensions expressly defined in the offer. First, income took one of seven values. The middle value was 19.00 USD per hour, roughly the median hourly wage for an administrative assistant in the US.¹¹ We include three values above and below, each a step of 0.50 USD. The top and bottom values approximate the 25th and 75th percentile of pay for this category of work in the US. Second, independence varied with the inclusion of either the statement ‘You will need my approval about *once* a week’ or ‘You will need my approval about *twice* a week’. (We avoid the use of imprecise language such as “high”, “medium” or “low”. [Johnston et al., 2017]) Third, potential for advancement is varied by inclusion of either the statement ‘If everything goes well, I could see you move up in about *one* year’s time’ or ‘If everything goes well, I could see you move up in about *two* year’s time’. These elements may be of incidental interest in their own right (such as in Eriksson and Kristensen [2014]), but their inclusion here is primarily for obfuscation, to shroud the true variables of interest. The design reduces respondents’ psychic discomfort in expressing prejudice - there is a sort of plausible deniability in picking a job offered by a white in favor of a black manager if the jobs differ in other dimensions simultaneously [Fisher, 1993].

Our interest is in how the attractiveness of a job varies with the race and gender of the person offering the job. Race and gender is introduced

¹¹<http://www1.salary.com/Administrative-Assistant-I-Salaries.html>

into our job offers by the name and photograph contained in the signature at the bottom of each offer. Each offer is made by a prospective manager from one of four types - a white male, white female, black male or black female. The selection of names and pictures will be discussed below. The controlled variation of these elements delivers the treatments in our research design.

The prescribed job characteristics allow $(2 \times 2 \times 7 \times 4) = 112$ distinct offers (and 12,544 distinct offer pairs) to be generated. For 16 of the 20 pairs faced by a respondent, elements are selected randomly. (In other words, respondents are facing differently configured job offers.) In terms of the race and gender of the person offering the job, a typical respondent would expect to see most of the possible permutations.¹² The responses to these questions are the basis for our main estimations.

In choice experiments it is standard practice to include tasks that are common to all respondents, often called fixed tasks. We include 4 such pairs which appear at the same points in the series for each respondent. In each, only the characteristics of the offerer and the hourly wage vary. In fixed task 1, a black male manager offers 0.50 USD more per hour than a white male manager. Fixed task 2 features two white female managers, one offering 3.00 USD more per hour than the other. This serves as an attention check. In fixed task 3, a white male manager offers 0.50 USD more per hour than a black male manager. In fixed task 4, a black female manager offers 0.50 USD more than a white female manager. In each of the fixed tasks, the left-right position of the wage advantaged job offer was randomized (in the respondent's browser). For every respondent the fixed tasks appeared at positions 4, 8, 12 and 16 in the 20 task sequence.

To ensure sufficient power we recruited a sample size of about 1000 respondents, each deciding between 16 randomly-generated and 4 common job offer

¹²Since each offer is randomly assigned a face, and there are four types of face (white male, white female, black male, black female) there are sixteen different permutations that might come up in any randomized pair. In a large sample we expect the proportion of job offers from white managers (ignoring gender) to converge to 0.5, and so on.

pairs.¹³ We restrict our analysis to white and black respondents.¹⁴ The now mutually exclusive respondent groups sum to 909 respondents. The 14,544 idiosyncratic choices of our sample places us in the top sample size decile of those reviewed by de Bekker-Grob et al. [2015]. Other choice experiments centered on labor markets use 21,658 choices [Eriksson and Kristensen, 2014], 4,112 choices [Wiswall and Zafar, 2016] and 3,245 choices [Mas and Pallais, 2017].

Names and Pictures

Gender and race are conveyed by the combination of a name and photograph at the bottom of a job offer.

For each of our four race \times gender treatments (white male, white female, black male, black female) we constructed a pool of 20 names. We combine first names from the ‘Blackest and Whitest Names in America’ list [Levitt and Dubner, 2011] with the blackest and whitest surnames from the 2010 U.S. Census. A typically black male name constructed in this way is Tyrone Robinson, whereas a typically white female name is Emma Reilly. The full list of 4×20 names created in this way is presented in Table 1. When required, names are drawn at random from the appropriate pool. (For example, when a white male name is needed one of the names from the first column of Table 1 is randomly drawn.)

For each of our race \times gender treatments we also construct a pool of 20 photographs from the University of Chicago Face Database (CFD).¹⁵ The CFD

¹³One often cited rule in this literature is Orme [1998] who postulates that sample size N for a choice experiment should satisfy

$$N > \frac{500 \times c}{t \times a} \quad (2.1)$$

where t is number of tasks, a is number of alternatives per task and c is number of ‘analysis cells’. Analysis cells are either the largest number of levels for an attribute, or the largest interaction. In our study when estimating main effects this rule recommends a minimum of 110 respondents. When testing first order interactions it implies a minimum of 219 respondents. We are far above these recommended floors.

¹⁴We excluded the tiny number of respondents who indicated that they considered themselves black *and* white or male *and* female.

¹⁵<http://faculty.chicagobooth.edu/bernd.wittenbrink/cfd/index.html>. There is a broader and interesting set of papers that investigate the role of faces. Eckel and Petrie [2011] and Heyes and List [2016] use faces in a laboratory setting, whereas Charness and Gneezy [2008] uses surnames. Rule and Ambady [2008] connect CEO appearance to company profits.

provides a large number of high resolution photographs of male and female faces of different races, designed for research purposes. The photographs are standardized including subject clothing, lighting, background and expression. Extensive norming data are provided for each photograph subject. These include both objective (e.g. face size) and subjective characteristics (e.g. attractiveness, trustworthiness, etc.) as rated by a panel of Chicago-area residents.¹⁶ For a detailed description of the CFD, tests of its validity and discussion of appropriate applications, see Ma et al. [2015].

Faces vary in more dimensions than race and gender. For this analysis we attempted to decant out as much of the non race and gender variation in face pools as possible. In other words, to avoid having pools of faces that are significantly different *as a set* in terms of characteristics that might plausibly influence job choice. For example, we would not want the 20 white male faces used in the experiment to be systematically more trustworthy-looking than the 20 black male faces, or the 40 white subjects to be systematically younger than the 40 black subjects.

To this end, we randomly drew four samples of twenty photographs from the larger CFD population; one for each of the four race \times gender treatments. Observing the CFD ratings along the ‘subjective’ dimensions, we use standard multivariate analysis of variance to test whether the black and white photographs in the samples differed significantly along (a) any individual attribute or, (b) across attributes collectively. For (b) we report only Hotelling’s R-squared, as when comparing two groups the most common MANOVA measures collapse to it. We test whether the ‘cloud’ of photos for one treatment (notionally plotted in the 17-dimension characteristic space) is not significantly different from the cloud of another treatment. For race, we test for differences between the 40 white and 40 black photographs, between the 20 white and black males, and between the 20 white and black females. If a statistically significant difference was detected *either* in a single dimension or collectively, we replaced and redrew all samples afresh and tested again. We repeated this exercise until we arrived at pools of photographs that satisfied our requirements. The resulting pools of faces are displayed in Figure 2. Table A3 presents the insignificant differences between white and black photographs.¹⁷

¹⁶The attributes rated subjectively are: how old, afraid, angry, attractive, baby-faced, disgusted, dominant, feminine, happy, masculine, prototypical, sad, suitable for research, surprised, threatening, trustworthy and prototypical the subject appeared.

¹⁷Photographs were not balanced across genders in the same manner. Table A4 shows that the male pool of faces were significantly (a) more dominant, (b) more threatening and

Faces are drawn at random from the appropriate pools and inserted into the job offers. For example, when a white male face was needed by the design, a photograph from the top two rows is randomly selected and displayed.

2.2.3 Data Manipulation

For the purposes of analysis we approach the data in three different ways.

First, by looking at three of the fixed questions (excluding one designed to act as an attention check) we observe the propensity among different respondents to directly trade race and income. Note that a particular subject choosing a job offered by a white over one offered by a black does not in itself imply race-based preferences. Individual pairs of faces vary in attributes other than race. However, across a large set of decisions the randomization of faces across respondents makes systematic patterns of choice in favor of one race or another plausibly indicative of such preferences at the group level.

Second, using the choices made by subjects in the 16 randomly-generated job offer pairs, we observe aggregate ratios of white offers chosen to black offers chosen and male offers chosen to female offers chosen between respondent types. For example, amongst all 5,184 pairwise choices made by the 324 white male respondents, we observe how many choices were made in favor of white managers and how many in favor of black managers and calculate the ratio. Given the orthogonality of job offer elements, we expect the ratio to be one, absent race preferences.

Third, we use the choices made by respondents in the 16 random pairs to estimate acceptance probability elasticities to the characteristics of job offers. With these elasticities, we calculate race and gender WTPs at the group and individual respondent levels.

The natural framework to estimate acceptance probabilities is the random utility model. Decision makers choose among a set of alternatives, and we observe those decisions. We assume that decision makers are utility maximizers who choose the alternative that affords them the greatest utility. Although we do not observe the decision maker's utility, we do observe the attributes of the alternatives. We allow some part of the decision maker's utility to be random, and consider the probability of an alternative being chosen as a function of the desirability of its attributes.¹⁸ Over repeated choices, we quantify how deci-

(c) less trustworthy than those in the female pool. These differences plausibly push results *in favor* of offers from females making any bias towards female over male offers understated.

¹⁸Different functions connecting decision utility to choice probability lead to different

sion makers value attributes. In our study, workers are given a choice between two job offers. We assume that they choose the offer that provides them with the greatest utility. For our initial analysis, we apply the multinomial logit model (MNL) [McFadden, 1974] to estimate acceptance probability elasticities and group-level WTPs.¹⁹ We then produce distributions of individual-level WTPs to explore how they vary with *respondent* characteristics.

Our WTP measures closely resemble the hedonic framework also widely applied by economists. The hedonic compensation model [Rosen, 1974, 1986] is a well established framework studying the labor supply to occupations that are differentiated by wage and job amenities. In essence, jobs with favorable characteristics attract labor at lower wages whereas those with unfavorable characteristics require higher wages from a compensating differential. Our experiment confirms intuition that faster advancement, greater independence and higher pay are all characteristics that are attractive to job-seekers. Interpreted within this framework, we investigate how the race and gender of a manager affects the favorability of the proposed working conditions. Do workers need to be compensated if the manager is female? Or are workers compensated with an own-race manager such that they would forego higher earnings available elsewhere?

2.3 Results

2.3.1 Sample

Responses were restricted to MT workers located in the United States who had never previously completed work for the requesting account. Unique response was ensured using log-in ID and browser blocking. We present results for those who completed the entire experiment (4.6% failed to finish).

MT workers choose which Human Intelligence Tasks to complete, making ours a convenience sample. Luckily, our focus on race-gender subject pools relieves the need to collect a sample representative in terms of race and gender. Summary statistics, decomposed into our four race-gender pairs are presented in Table A2. In terms of differences, black respondents are an average of 2.6

functional forms, such as logit and probit.

¹⁹The MNL embeds particular assumptions. To challenge robustness we also conduct parallel analyses using the main competitor models, multinomial probit (MNP) and mixed logit (MXL), which embed different assumptions. Results are largely unchanged. [Cameron and Trivedi, 2005, p.472]

years younger than white respondents. All other differences on observables are statistically insignificant between white and black respondents. For males, blacks and whites have no statistically significant differences. The black females in our sample are 3.78 years younger and less likely to be single and more likely to be separated than their white counterparts. We later confirm that including respondent demographics (other than race and gender) as controls does not substantially affect results.

In the rest of this section we present the results of the three approaches to the data outlined earlier.

2.3.2 Approach 1: Fixed Tasks

Recall that four tasks were fixed for every respondent. They appeared among other tasks for purposes of camouflage. By design, they required respondents to directly trade rate of pay and the race of offer manager.²⁰ The structure of the tasks and response patterns are presented in Table 2.

Fixed task 1 paired otherwise identical job offers. A black male offering an hourly wage 0.50 USD higher than a white male manager. A quarter (25.0%) of white male respondents, and 23.0% of white females, accepted the lower paid white offer. Only 15.7% of black males and 1.2% of black females accepted the white manager's offer.

Fixed task 2 acts as an attention check. It paired two job offers that were identical except the rate of pay for one was higher by 3.00 USD per hour.²¹ We interpret any choice of the lower paid job to indicate inattention. Encouragingly, few respondents fail the attention check (2.3% overall). The rate of failure was slightly higher among black respondents than whites. As a robustness check, we demonstrate that excluding those that failed our attention check has no substantive effect on results.

Fixed task 3 mirrored fixed task 1 but with races reversed. A white male offering an hourly wage 0.50 USD higher than a black male manager. Only 10.8% of white male respondents accepted the lower paid black offer, (8.6%

²⁰The fixed tasks were common qualitatively across subjects, but were not identical. While the design might prescribe a black female face, the precise black female face is randomized. Since faces vary in characteristics other than race and gender, a non-trivial preference between faces does not necessarily imply that any race or gender preference at the level of an individual.

²¹Fixed task 2 was uniquely programmed to present two offers from the *same* photographed and named manager

of white females). A large share of black respondents choose the lower paid black offer - 45.1% of males and 44.4% of females.

Fixed task 4 paired otherwise identical job offers. A black female offering an hourly wage 0.50 USD higher than a white female manager. The results are similar to fixed task 1. 20.4% of white male respondents accept the lower paid white female offer (18.3% for white female respondents). Only 11.8% of black male respondents choose the lower pay white offer, (8.6% for black females).

Taken together, the responses to the fixed questions are suggestive of a preference for own-race job offers, further investigated later. For now, we apply a simply linear probability model (LPM) to the choices made by respondents in fixed task 1. The dependent variable is binary and takes the value 1 if the respondent chooses the lower pay/white offer and 0 otherwise. The regressors of interest are respondent characteristics (offer attributes are otherwise fixed). The results of this exercise are summarized in Table 3, with specifications less sparse from left to right. In the most general specification (column 5) the respondent being white increases the probability of choosing the lower pay/white offer by 16.3%.²² Gender alone has no significant effect and the white \times male interaction term is insignificant at conventional levels. Column 6 confirms that re-estimating the preferred specification while excluding respondents who failed the attention check (21 out of 909 = 2.3%) makes no discernible difference to the coefficient of interest.

Table 4 reports the results of conducting the same analysis on responses to fixed task 3. Here, the dependent variable takes value 1 if the respondent chooses the lower pay/black offer and 0 otherwise. In the most general specification (column 5) the respondent being white decreases the probability of choosing the lower pay/black offer by 34.3%. Gender alone has no significant effect and the white \times male interaction term is insignificant at conventional levels. Dropping those who failed the attention check does not disturb results.

Table 5 reports corresponding results relating to fixed task 4. Here the dependent variable takes value 1 if the respondent chooses the lower pay/white female offer instead of the higher pay/black female offer and 0 otherwise. In the most general specification (column 5) the respondent being white increases

²²In the post-experiment survey respondents were asked to indicate how important they regarded the various dimensions of a job. In particular they ranked pay level, independence, prospects for advancement and what their manager ‘seemed’ like from most to least important. These are contained in the ‘job priority controls’ in most tables. Summary statistics presented in Table A2. We will see that the inclusion or exclusion of these as controls has little impact on results.

the probability of choosing the lower pay/white offer by 9.4%. Gender alone has no significant effect and the white \times male interaction term is insignificant at conventional levels. Dropping those who failed the attention check does not disturb results.

Table A5 (in appendix) reports results for the same analysis conducted on fixed task 2 which served as the attention check question. We find no systematic effect of race or gender on inattention.

2.3.3 Approach 2: Aggregate Choice Patterns

The second way we approach the data exploits in aggregate the choices made by respondents to 16 idiosyncratic tasks.²³ For these tasks, each element of every job offer was randomized from the available values. This included the race and gender of the person making the offer, but also the wage, measure of work independence and speed of advancement. If respondent choice behavior is *not* race-sensitive, then as the sample of responses gets large the probability that a job offer from a black photograph is accepted should converge to the complementary probability that a job offer from a white manager is accepted. Equivalently, the ratio of probabilities should converge to unity. The same applies for male and female offers.²⁴

Table 6 reports these probabilities for each race-pair type of respondent. The first numbers in column 1 of Table report that amongst the 5,184 randomly generated job offer pairs submitted to white males, these respondents were 10.7% more likely to accept a job offered by a white than an otherwise equivalent job offered by a black.²⁵ This implies a statistically significant (at 1%) race-sensitivity of choices for white male respondents. Similarly, white females favor white offers by 11.7% in their choices, with the bias again significant at 1%. Black respondents also display own-race preference, accepting job offers from white managers 9.3% (males) and 13.9% (females) less often.

Table 6 also reports the results for the specious elements of the job offer. The results are intuitive. In the universe of responses, probability of offer acceptance is monotonically increasing in wage rate. All race-gender respon-

²³To preserve equality in how often attribute levels are presented, we exclude the 4 fixed tasks that were common across respondents.

²⁴Indifference in an attribute would lead to a proportion for each of 0.50 and therefore a ratio of 1.00. Alternatively, if respondents were evenly split between those who prefer A and those who prefer B we would expect the ratio to converge on 1.00.

²⁵Equivalence here should be understood in a stochastic sense. All other elements of the jobs offered in any particular task are randomized.

dent groups choose more independence and (most pronouncedly) more rapid advancement.

2.3.4 Approach 3: Probabilistic Choice Modeling

Our third approach is to apply the multinomial logit model (MNL) following McFadden [1974] and others. In a discrete choice setting, MNL estimates how the likelihood a particular alternative is chosen changes with the attributes of the alternatives. How often an alternative with a particular attribute is chosen over repeated decisions is connected to a measure of its desirability. For ease of interpretation we report acceptance probability elasticities (the marginal effects of the model) and WTPs for job attributes, rather than likelihoods.

The results of this exercise, conducted on each race-gender respondent type separately, are summarized in Table 7. In the top row, we observe that replacing a black with a white photograph in the email signature increases the probability of acceptance by a white male respondent by 11.3%. For white female respondents that number is 7.9%. The same change makes a black male respondent 9.6% *less* likely to accept and black females 17.3% *less* likely to accept. Each of these coefficients achieves significance at a level better than 5%.²⁶

Observed choices are much less sensitive to gender of offer. While within each respondent group the estimated coefficient on male is negative (consistent with a bias toward choice of job offers from females) the coefficients are three to ten times smaller in absolute value than for race. Only white female respondents ever achieve statistical significance for gender preferences.

The bottom rows of Table 7 confirm our intuition for the non race/gender job offer attributes. For every type of respondent, a job offer is more likely to be accepted if the wage is higher, anticipated advancement faster and independence greater. There are some variations in responsiveness - for example females (both black and white) are substantially more sensitive to independence in the workplace.

²⁶The models without interaction terms are presented throughout. When all two-way interaction terms are included, estimates remain stable across respondent subgroups, particularly for race and gender.

WTP Estimates - Group and Individual

Our analysis estimated the acceptance probability elasticities to job offer attributes. Combining the non-pecuniary and wage elasticities produces an implied WTP for attributes in terms of foregone wage. If the WTPs for manager attributes are non-zero, respondents are willing to forego some part of wages ‘in exchange for’ a manager of preferred race-gender type.

For each respondent type, we divide the logit model coefficient for a job attribute by two times the estimated coefficient for income (wages per hour was presented in 0.50 USD steps). The implied WTP estimates are presented in Table 8. A positive value indicates respondents have a positive WTP (will accept a lower wage) for an offer with that attribute. A negative value indicates respondents have a negative WTP and must be compensated with additional wages to accept an offer with that attribute.

For each respondent type we find a non-zero WTP to replace a black offer by a white offer, in every case with statistical significance at better than 1%. The WTP for a white manager is positive for white respondents and negative for black respondents. The coefficients can be interpreted directly as USD wages per hour. For example, white male respondents would accept around 0.31 USD less per hour in wages to replace a black with a white manager. Assuming a work year of 2000 hours, this corresponds to an annualized value of 624 USD. Black female respondents have the largest WTP, as they are willing to pay 0.39 USD per hour to have a black rather than white manager (788 USD per annum).

The WTP for a male rather than female manager can be interpreted similarly. While each WTP is negative (consistent with a universal preference in favor of an offer from a female) the largest coefficient value, and the only one that obtains significance at better than 10%, is from white female respondents. However, the absolute value of 0.07 USD per hour is small.²⁷ We have so far have analyzed group level choice patterns, and the sensitivity of choices to job offer attributes. This is a common approach in choice modeling exercises. However, using either the 16 or the 20 (by including fixed tasks) pairwise choices made by each individual respondent, we estimate *individual* WTPs in the same manner. We then explore how those individual-level measures are distributed for each respondent group. We are not interested in the choice

²⁷Incidental to our focus we can observe that each respondent group has a substantial WTP for (a) greater independence and (b) more rapid advancement (each group is willing to forego around one dollar per hour to accelerate anticipated promotion by 1 year).

patterns of any single respondent, only how they vary in the larger samples. Because of the small number of data points for each individual respondent in many cases the point estimated individual WTPs do not achieve statistical significance.

The results of this exercise are individual-level WTPs for white managers. Histograms are displayed in Figure 3 (with the distributions winsorized at 5%). The top left panel presents WTP estimates for the 324 white male respondents, the top right for the 453 white female respondents, and so on. Consistent with results already presented, there is significant WTP heterogeneity between respondent types. The two upper (white respondent) panels exhibit a clear rightward mass shift (preferring white offers). The two bottom panels, which depict WTP histograms for black respondents, exhibit a leftward mass shift (preferring black offers).

Overall, this approach illuminates *within*-type variation. We estimate a positive WTP for a white manager for 63.0% of white male respondents (63.8% of white females). Conversely, we estimate a positive WTP for a black manager for 72.5% of black males (82.7% of black females).

What is clear is that results are *not* driven by outlier respondents, but rather reflect popular preferences of our respondents. The exception to this is among black female respondents, where we can see that the winsorization had more ‘bite’ and 19.8% of respondents showed a substantial (greater than or equal to 1.25 USD) WTP to avoid a white manager.

We use the fixed task results from earlier to benchmark our individual level WTPs. For fixed task 1, 25.0% of white male respondents revealed they were willing to accept 0.50 USD less in exchange for a white manager. Our individual-level estimates predict that 22.2% of white males have a WTP for a white manager exceeding 0.50 USD. 23.0% of white females chose a white manager in fixed task 1, our individual-level estimates indicate 23.4% of white females would make the same decision. For black males we estimate 11.8% with a WTP for a white manager exceeding 0.50 USD, and for black females this number is 2.5%.

To formalize what ‘eyeball’ analysis of Figure 3 already reveals, we use OLS to characterize the association between our individual-level WTP estimates with the race and gender of respondent. The results of this exercise are summarized in Tables 9 and 10. In each case, the preferred specification is presented in column 5, including individual socioeconomic demographics and stated job preferences as controls, and winsorizing to control for extreme

tastes.²⁸ Inspecting Table 9, individual WTP to replace a black manager by a white is 0.64 USD higher for a white respondent than a black respondent, and only a further 0.001 USD if the respondent is additionally male. This is consistent with the group level results already seen. Column 5 in Table 10 confirms that WTP for a male compared to female manager is not discernibly affected by the race or gender of respondent.

Robustness and Some Additional Exercises

In this section we challenge the robustness of our model, and report the results of some supplementary exercises.

Alternative estimation The MNL specification adopted for our primary results assumes that estimating errors are independent and identically distributed in the random utility model. For the purposes of robustness, we re-estimate the acceptance probabilities, group WTPs and individual level WTPs using the multinomial probit (MNP) model, which assume only that the error terms are jointly normally distributed.²⁹

We present the probability acceptance elasticities in Appendix Table 6 (which corresponds to Table 7), now estimated by the probit function. The sign and significance of all coefficients of interest is sustained, while all coefficients (including the specious ones) are slightly smaller in absolute value. The implied WTP estimates are reported in Appendix Table 7, which are negligibly different from those presented in Table 8.

We also re-estimate the WTPs using the mixed logit model (MXL), which does not require the normality assumption of the MNP. Further, MXL can estimate any discrete choice model following random utility maximization, making it a flexible and computationally practical approach to discrete choice analysis [McFadden and Train, 2000]. The WTPs that result are presented in Appendix Table 8 and are different only marginally from our preferred estimates (Table 8).

²⁸Demographics include respondent age, household income, marital status, educational attainment and labor force status. Stated Preferences are from the post-experiment survey. Respondents indicated what their most (and least) important aspect of a job offer was between income, independence, advancement and what their manager 'seemed' like. Statistics presented in Table A2.

²⁹Cameron and Trivedi [2005] mention there is little difference in practice between logit and probit models. Kropko [2007] runs Monte Carlo simulations showing that MNL provides better estimates than MNP, even when IIA is severely violated.

Confounding facial attributes The treatments of interest are the names and faces that are placed at the bottom of each offer mock email. Naturally, a particular photograph can vary in characteristics other than race and gender. If in a particular task a respondent rejects an offer from (say) a black face and accepts another otherwise identical one from a white, that choice cannot necessarily be said to have been due to the subject's race. It may be that the white face is perceived to be more honest, more friendly, or to dominate in some other dimension that the respondent favors.³⁰

These confounding facial attributes would invalidate inference, if correlated with race or gender. We sought to minimize such potential confounding. We sourced photographs from a stimulus set offered by the University of Chicago. Photos were standardized in facial expression, subject dress and image quality. Photographs in the database come with US survey-based subjective ratings on subjective 16 dimensions, what psychologists - including the curators of the CFD - refer to as norming data. From that larger population, we selected 4 pools of 20 photographs (one for each race-gender type). The pools were balanced between races, that is, not statistically different from each other along the rated dimensions either individually or in a multivariate sense. We were unable to convincingly balance between genders, but the photo pools applied plausibly biased things in favor of female over male offers. The fact that we do not use a single image to represent a race-gender type, but rather draw randomly from these larger pools, should also mean that results will only be minimally influenced by any single 'rogue' picture. As such it seems unlikely that such extraneous variation would seriously threaten our results.

An alternative approach, used in some other studies involving photographs of faces, would have been to use unbalanced pools of pictures and then use subjective ratings on pertinent directions as *controls* in regressions.³¹ Such an

³⁰Many studies provide convincing evidence that labor market outcomes favor people with faces that are subjectively more attractive. For example, Mobius and Rosenblat [2006] observe the different beauty-earnings channels using a lab experiment. Ruffle and Shtudiner [2014] find callbacks to attractive men are significantly higher in a field experiment. Hamermesh et al. [1994] show that labor outcomes are increasing in how good-looking subjects are rated in current photographs while Scholz and Sicinski [2015] find facial attractiveness of male high school graduates is correlated with future earnings. Todorov et al. [2005] find competence inferences from photographs predict outcomes (and victory margin) of US congressional elections. Chen et al. [2016] find that masculinity in lawyers reduces supreme court win probability.

³¹For example, Pope and Sydnor [2011] find that the race and gender of borrowers on the peer-to-peer lending site Prosper.com impacts their success in obtaining funds. Their main specification contains controls for three subjective elements (happiness, weight and

approach could easily fall victim to a bad control problem, and we do not favor it. Race or gender of subject might plausibly influence the rating of a picture on a non-race dimension (perhaps raters tend systematically to see black faces as more threatening, or female faces as more trustworthy).

As an additional falsification exercise we re-estimate our results (those in Table 8) but also include the ratings of individual faces along all 16 available dimensions (plus age). If we have balanced the photograph pools well, these additional controls should rarely, if ever, be significant. This turns out to be the case. Most importantly, they do not disturb our race WTP estimates. Comparison of the estimated coefficients across the top rows of Tables 8 and 11 reveals that the inclusion of the suite of controls for confounding facial attributes has little impact on our race conclusions. Significance is lost on the positive WTP for a female offer among white-female respondents, consistent with our observation that the photo pools were likely biased in favor of female offers.³² It is also evident that preferences for female bosses in general may have been due to photo characteristics, as white male and black males now uniquely exhibit a positive WTP for a male manager, albeit they are statistically insignificant. This leads us to suspect that our photo balancing act was a necessary step in distilling the effect of race from other photo characteristics.

Attention A small number of respondents (2.3% of the total) failed the attention check presented in fixed task 2. Following Mas and Pallais [2017] we used a dominated job offer to try to remove the effect of inattentive respondents. Since we only provide two options and the respondent is forced to choose one of them this is only a coarse check, 50% of respondents making selections at random would be expected to choose the ‘correct’ option by chance.

Appendix Table 9 reports the result of re-estimating our primary analysis (reported in Table 8) on the sample of responses from respondents who did not fail the attention check. Sign and significance of estimates is sustained in every case. Coefficient sizes are little disturbed.³³

attractiveness, each on a three point scale).

³²For example, female photographs were more trustworthy - white female respondents may have been responding to the trustworthiness of photographs. We were unable sensibly to construct photo pools that eliminated the imbalance: The CFD raters consistently identify females as looking more trustworthy than males.

³³Also relating to attention at the group level we monitored the proportion of time respondents chose the offer presented on the left and right side of the screen. A strategy of “clicking through” the exercise to get to the end might be expected to lead to repeated

Active job-seekers The sample used for our primary estimates are a convenience sample of MT workers who elected to complete our HIT. However, of particular interest might be the choice patterns of those actively involved (recently or prospectively) in real-world job offer settings. We are able to address this in two different ways.³⁴

In the post-task questionnaire 60.5% of respondents declared themselves ‘likely’ or ‘very likely’ to apply for a job in the next 12 months. The results of re-estimating our central specification on this sub-sample are summarized in Table 12. The sign of the estimated coefficient is sustained in each of the four race-gender respondent types. Only black male responses lose significance (the coefficient is little changed but sample size is reduced by about one-third).

We also re-estimate WTPs for the sub-sample (52.7%) of respondents who answered yes to the question ‘Have you applied for a job in the previous 12 months?’ The results are summarized in Table 13. Again, results remain similar to the those derived from the overall sample.

Fatigue Respondents face a series of tasks and it is possible that the patterns of responses evolve over repeated tasks (e.g. changing with increasing fatigue or decreasing attention), threatening inference.

We test the stability of responses following Johnson and Orme [1996] by splitting the 16 randomized questions faced by each respondent into the first 8 and the second 8 in the task series and running our primary specification on the two halves separately.³⁵

The results of this exercise are reported in Table A10. Qualitatively the pattern of choices remains fairly consistent between the top and bottom panels. The only eye-catching change is the increased own-gender bias that is revealed by white female respondents *only* in the second half of the sample.

selection of the left hand or right handed offer [Krosnick, 1991]. We are unable to reject the hypothesis that total left and right choices in the sample are binomially distributed with $\pi = 0.5$ ($p = 0.50$).

³⁴The respondent summary statistics in Appendix Table 2 show that a large proportion of respondents in each subgroup are currently active in the labor market, employed either full-time, part-time or on contract. A smaller number are currently students or actively looking for work.

³⁵We have focused throughout on choice patterns rather than emphasize preferences, though the two are intimately linked under mild assumptions. In experiments involving repeated tasks, the expectation is that as a subject becomes fatigued from repetition he will tend to rely increasingly on heuristics to make decisions [Johnson and Orme, 1996, Boksem et al., 2005, Faber et al., 2012]. As such, our prior might be that choices made on tasks coming later in the series would tend to more closely reflect innate or true preferences.

Following Johnston et al. [2017], we also run the analysis again using only the first question for each respondent, to test for "sequencing" effects. White males retain the same WTP for a white manager. Black females have a more extreme preference for a black manager. White females and black males are not statistically different from 0. The latter is likely due to sample size, whereas the former indicates that white females require multiple offers before expressing race based tastes. Results available upon request.

2.4 Conclusions

Extensive anecdotal and empirical evidence points to how demand side decisions in the labor market are sensitive to the race and gender of workers. Our research objective was to cast light on the obverse, to ask whether the *supply* side of the labor market favors employers of a particular race or gender. We believe we are the first to address this question.

In a laboratory-in-the-field experiment conducted in the online marketplace Mechanical Turk we found persuasive evidence that not only does such favoritism exist, at least with regard to race, but is quite broad-based. Respondents were asked to choose between a series of hypothetical job offer pairs that varied in several substantive regards. Our focus was on how choice patterns were sensitive to the race and gender of the person making the offer. As such we manipulated the name and thumbnail pictures placed in the signature strips at the bottom of each e-mail of offer.

Approaching the responses in a number of different ways we found a consistent and pronounced bias in choice patterns towards own-race offers. There was little consistent evidence of bias on the basis of gender, with the exception of a small (but statistically significant) preference for white female offers among white female respondents. Results proved remarkably robust.

The study is a first step in the direction of understanding a previously unstudied phenomenon. We have been careful not to over-interpret the results. Though earlier research points to the personal values and responses on social and political questions derived from exercises on MT being very similar to those derived from parallel studies on population-representative platforms, extrapolation from our sample to any wider population requires caution. Further experimental and empirical work can probe further questions of external validity.

What are the implications if the qualitative results are sustained? We already know that facial characteristics play a significant role in strategic set-

tings [Eckel and Petrie, 2011, Heyes and List, 2016]. If a manager of particular race-gender type finds it harder to recruit staff than a peer of a different type, then it is straight-forward to predict that this could impact the ability of an individual of that type to flourish in managerial roles and provide an additional challenge to their promotability in a competitive market for managers.³⁶

³⁶Our objective in this chapter has not been to spell-out the social consequences of the choice bias of which we find evidence. It provides another mechanism for transmission of biased (discriminatory) preferences of one group to outcomes for another. If in a product or service market customers prefer to be served by an agent of a particular type then competition will cause that type of agent to be favored (put crudely, if restaurant clients in a particular locale prefer not to be served by black waiters, then competitive pressure will lead restaurant owners to hire white waiting staff). Similarly, if a black manager must pay more than a white manager to fill a subordinate position, and the pool of potential subordinates is white, we would expect to see disproportionately more whites in managerial roles.

2.5 Tables and Figures

Table 1: Manager Names

White Male	White Female	Black Male	Black Female
Hunter Brennan	Emma Reilly	Deshawn Jefferson	Imani Washington
Jake Fischer	Allison Gallagher	Deandre Banks	Ebony Booker
Wyatt Mueller	Claire Schmitt	Marquis Mosley	Shanice Joseph
Cody Novak	Emily Weiss	Darnell Charles	Aaliyah Jackson
Dustin Klein	Katie O’Connell	Terrell Rivers	Precious Dorsey
Luke Schneider	Madeline Schroeder	Malik Mack	Jazmine Mays
Jack Koch	Katelyn Kramer	Trevon Williams	Deja Singleton
Scott Huber	Molly Yoder	Tyrone Robinson	Diamond Branch
Logan Bauer	Abigail Schmidt	Willie Coleman	Jazmin James
Cole Erickson	Carly O’Donnell	Dominique Harris	Aliyah Roberson
Lucas Becker	Jenna Schaefer	Demetrius Benjamin	Jada Glover
Bradley Meyer	Heather Roth	Reginald Hinton	Tierra Houston
Jacob Schultz	Katherine Carlson	Jamal Hampton	Tiara Clay
Garrett Olson	Caitlin Larson	Maurice Sims	Kiara Flowers
Dylan Berg	Kaitlin Knapp	Jalen Wiggins	Nia Gaines
Maxwell Hess	Holly Jacobson	Darius Terrell	Jasmin Dixon
Connor Friedman	Amy Krueger	Xavier Franklin	Asia Ware
Brett Walsh	Kaitlyn Rasmussen	Terrance Daniels	Jasmine Thomas
Colin Stein	Hannah Christensen	Andre Tate	Alexus Jones
Tanner Schwartz	Kathryn Dougherty	Darryl Randolph	Raven Grant

Notes: Given names are drawn from the ‘Blackest and Whitest Names in America’ list. Levitt and Dubner [2011]. Surnames drawn from the 2010 US census in the same manner. Names are drawn randomly when questions are generated.

Table 2: Fixed Task Responses

Task	Manager Race	Manager Gender	Wage Offered	Respondent Type			
				White		Black	
				Male %	Female %	Male %	Female %
Fixed 1	White	Male	\$19.00	25.00	22.96	15.69	01.23
	Black	Male	\$19.50	75.00	77.04	84.31	98.77
Fixed 2	White	Female	\$20.50	97.22	98.01	94.12	100.00
	White	Female	\$17.50	02.78	01.99	05.88	00.00
Fixed 3	White	Male	\$19.50	89.20	91.39	54.90	55.56
	Black	Male	\$19.00	10.80	08.61	45.10	44.44
Fixed 4	White	Female	\$19.00	20.37	18.32	11.76	08.64
	Black	Female	\$19.50	79.63	81.68	88.24	91.36
Respondents				324	453	51	81

Notes: Every respondent was shown the above four tasks. In each, the level of independence, advancement and gender between job offers was the same. Fixed task 1 displayed a black male manager offering 0.50 USD more than a white male manager, which 75.00% of white males accepted.

Table 3: Fixed Task 1 - Black Male Offer > White Male Offer

	1	2	3	4	5	6
White	0.170*** (0.038)		0.217*** (0.049)	0.218*** (0.048)	0.211*** (0.048)	0.209*** (0.048)
Male		0.041 (0.028)	0.145** (0.072)	0.121* (0.071)	0.127* (0.072)	0.081 (0.073)
White * Male			-0.124 (0.078)	-0.112 (0.077)	-0.115 (0.078)	-0.080 (0.078)
Constant	0.068* (0.035)	0.197*** (0.018)	0.012 (0.045)	0.246** (0.112)	-0.010 (0.179)	-0.072 (0.189)
Respondents	909	909	909	909	909	888
Job Priority Controls				YES	YES	YES
Demographic Controls					YES	YES
Attention Check Passed						YES

Notes: Depicted are coefficients of a linear probability model with dependent variable indicating a respondent's choosing the lower wage and white offer. Fixed task 1 presented respondents with a white male manager offering 0.50 USD less than a black male manager, all other attributes equal. Column 1 indicates that a white respondent is 17.0% more likely to choose a white manager with a lower wage. Respondents that indicated they are both white and black removed. Respondents stated their job attribute preferences at the end of the experiment. Demographic controls include marital status, education and employment. Standard errors in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table 4: Fixed Task 3 - White Male Offer > Black Male Offer

	1	2	3	4	5	6
White	-0.352*** (0.031)		-0.358*** (0.040)	-0.359*** (0.040)	-0.348*** (0.040)	-0.358*** (0.039)
Male		0.014 (0.024)	0.007 (0.059)	-0.016 (0.059)	-0.013 (0.060)	0.005 (0.060)
White * Male			0.015 (0.064)	0.032 (0.064)	0.017 (0.064)	0.002 (0.064)
Constant	0.447*** (0.029)	0.140*** (0.015)	0.444*** (0.037)	0.604*** (0.093)	0.750*** (0.148)	0.603*** (0.155)
Respondents	909	909	909	909	909	888
Job Priority Controls				YES	YES	YES
Demographic Controls					YES	YES
Attention Check Passed						YES

Notes: Depicted are coefficients of a linear probability model with dependent variable indicating a respondent's choosing the lower wage and black offer. Fixed task 3 presented respondents with a white male manager offering 0.50 USD more than a black male manager, all other attributes equal. Column 1 indicates that a white respondent is 35.2% less likely to choose a black manager with a lower wage. Respondents that indicated they are both white and black removed. Respondents stated their job attribute preferences at the end of the experiment. Demographic controls include marital status, education and employment. Standard errors in parenthesis. *** p<0.01, **p<0.05, * p<0.10

Table 5: Fixed Task 4 - Black Female Offer > White Female Offer

	1	2	3	4	5	6
White	0.093*** (0.036)		0.097** (0.046)	0.096** (0.045)	0.094** (0.045)	0.091** (0.045)
Male		0.023 (0.026)	0.031 (0.068)	0.009 (0.066)	0.000 (0.068)	0.015 (0.068)
White * Male			-0.011 (0.074)	0.000 (0.072)	0.005 (0.073)	-0.012 (0.073)
Constant	0.098*** (0.033)	0.169*** (0.017)	0.086** (0.042)	0.630*** (0.105)	0.431** (0.167)	0.444** (0.178)
Respondents	909	909	909	909	909	888
Job Priority Controls				YES	YES	YES
Demographic Controls					YES	YES
Attention Check Passed						YES

Notes: Depicted are coefficients of a linear probability model with dependent variable indicating a respondent's choosing the lower wage and white offer. Fixed task 4 presented respondents with a white female manager offering 0.50 USD less than a black female manager, all other attributes equal. Column 1 indicates that a white respondent is 9.3% more likely to choose a white offer with a lower wage. Respondents that indicated they are both white and black removed. Respondents stated their job attribute preferences at the end of the experiment. Demographic controls include marital status, education and employment. Standard errors in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table 6: Idiosyncratic Task Responses

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
Manager Race				
<i>White</i>	0.525	0.528	0.478	0.468
<i>Black</i>	0.475	0.472	0.522	0.532
<i>White\Black</i>	1.107***	1.117***	0.915	0.878**
Manager Gender				
<i>Male</i>	0.495	0.491	0.487	0.519
<i>Female</i>	0.505	0.509	0.513	0.481
<i>Male\Female</i>	0.981	0.966	0.947	1.080
Independence				
<i>Once a Week</i>	0.508	0.529	0.533	0.522
<i>Twice a Week</i>	0.492	0.471	0.467	0.478
<i>Once\Twice</i>	1.031	1.122***	1.142*	1.094
Advancement				
<i>One Year</i>	0.605	0.615	0.565	0.583
<i>Two Years</i>	0.395	0.385	0.435	0.417
<i>One\Two</i>	1.534***	1.597***	1.299***	1.400***
Income				
<i>\$20.50</i>	0.226	0.231	0.207	0.221
<i>\$20.00</i>	0.204	0.202	0.184	0.211
<i>\$19.50</i>	0.168	0.183	0.199	0.184
<i>\$19.00</i>	0.145	0.142	0.154	0.137
<i>\$18.50</i>	0.118	0.111	0.120	0.110
<i>\$18.00</i>	0.079	0.082	0.070	0.085
<i>\$17.50</i>	0.060	0.049	0.066	0.053
<i>Avg. Ratio</i>	1.255***	1.308***	1.233***	1.278***
Respondents	324	453	51	81
Responses	5184	7248	816	1296

Notes: Depicted are the observed choices patterns. White males accepted offers from white managers 52.5% of the time (White males are 10.7% more likely to accept a white offer compared to a black offer.) Statistical significance are proportion tests with $H_0 : p = 0.5$. Data are drawn from the 16 randomized questions. For an attribute that respondents have no preference over, we expect the proportion to be 0.50 or the ratio to be 1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table 7: Type Probabilities - Multinomial Logit

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.113*** (0.017)	0.079*** (0.015)	-0.096** (0.046)	-0.173*** (0.042)
Male Manager	-0.011 (0.013)	-0.030** (0.013)	-0.023 (0.030)	-0.016 (0.032)
Independence	0.059*** (0.017)	0.121*** (0.014)	0.080*** (0.033)	0.133*** (0.029)
Advancement	0.358*** (0.027)	0.444*** (0.025)	0.218*** (0.048)	0.473*** (0.059)
Income	0.181*** (0.012)	0.211*** (0.010)	0.132*** (0.022)	0.220*** (0.021)
Respondents	324	453	51	81
Responses	5184	7248	816	1296

Data are drawn from the 16 random questions. Coefficients are predicted changes in acceptance probability of a job offer. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Income is per 0.50 USD increase. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 8: Type WTPs - Multinomial Logit

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.312*** (0.047)	0.187*** (0.034)	-0.362*** (0.146)	-0.394*** (0.091)
Male Manager	-0.031 (0.037)	-0.072*** (0.031)	-0.087 (0.114)	-0.036 (0.073)
Independence	0.164*** (0.045)	0.287*** (0.032)	0.301*** (0.126)	0.303*** (0.066)
Advancement	0.987*** (0.064)	1.051*** (0.052)	0.823*** (0.180)	1.076*** (0.136)
Respondents	324	453	51	81
Responses	5184	7248	816	1296

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's logit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 9: Individual WTPs - White Manager

	1	2	3	4	5
White	0.836*** (0.171)		1.056*** (0.213)	1.014*** (0.216)	0.758*** (0.108)
Male		0.146 (0.124)	0.654** (0.321)	0.675** (0.328)	0.160 (0.164)
White * Male			-0.643* (0.346)	-0.642* (0.352)	-0.187 (0.176)
Constant	-0.492*** (0.158)	0.164** (0.079)	0.341 (0.542)	-0.505 (0.860)	-0.435 (0.430)
Respondents	886	886	886	886	886
Job Priority Controls			YES	YES	YES
Demographic Controls				YES	YES
Winsorized (5%)					YES

Notes: Depicted are OLS coefficients with dependent variable individual WTP for white manager. Column 1 indicates that a white respondent is willing to pay \$0.84 more than a black respondent for a white manager. Respondents that indicated they are both black and white, or male and female, removed. Respondents with an estimated negative utility of income removed. Respondents stated their job attribute preferences at the end of the experiment. Demographic controls include marital status, education and employment. Standard errors in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 10: Individual WTPs - Male Manager

	1	2	3	4	5
White	-0.022 (0.140)		0.135 (0.173)	0.119 (0.175)	0.097 (0.085)
Male		-0.063 (0.100)	0.226 (0.262)	0.167 (0.267)	-0.011 (0.130)
White * Male			-0.382 (0.282)	-0.327 (0.286)	-0.095 (0.140)
Constant	0.415*** (0.129)	0.422*** (0.064)	1.846*** (0.442)	2.055*** (0.699)	0.593* (0.341)
Respondents	886	886	886	886	886
Job Priority Controls			YES	YES	YES
Demographic Controls				YES	YES
Winsorized (5%)					YES

Notes: Depicted are OLS coefficients with dependent variable individual WTP for male manager. Column 1 indicates that a white respondent is willing to pay \$0.02 more than a black respondent for a male manager. Respondents that indicated they are both black and white, or male and female, removed. Respondents with an estimated negative utility of income removed. Respondents stated their job attribute preferences at the end of the experiment. Demographic controls include marital status, education and employment. Standard errors in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 11: Type WTPs - Photo Characteristics

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.287*** (0.050)	0.184*** (0.036)	-0.448*** (0.151)	-0.434*** (0.106)
Male Manager	0.118 (0.090)	-0.037 (0.077)	-0.084 (0.333)	0.060 (0.171)
Independence	0.166*** (0.045)	0.286*** (0.032)	0.315*** (0.122)	0.312*** (0.067)
Advancement	0.991*** (0.064)	1.050*** (0.052)	0.841*** (0.177)	1.075*** (0.135)
Responses	5184	7248	816	1296
Photo Characteristics	YES	YES	YES	YES

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's logit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 12: Type WTPs - Future Job Seekers

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.372*** (0.063)	0.217*** (0.044)	-0.197 (0.202)	-0.464*** (0.113)
Male Manager	-0.022 (0.053)	-0.080* (0.043)	-0.030 (0.159)	-0.079 (0.105)
Independence	0.114*** (0.059)	0.302*** (0.043)	0.377** (0.164)	0.274*** (0.101)
Advancement	1.043*** (0.088)	1.041*** (0.071)	0.787*** (0.219)	1.042*** (0.173)
Responses	3040	4368	592	800
Proportion	0.586	0.603	0.725	0.617

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's logit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 13: Type WTPs - Recent Job Seekers

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.299*** (0.061)	0.227*** (0.049)	-0.467* (0.233)	-0.392*** (0.127)
Male Manager	-0.019 (0.052)	-0.087* (0.047)	-0.056 (0.204)	-0.150 (0.117)
Independence	0.151*** (0.059)	0.271*** (0.047)	0.602*** (0.212)	0.223** (0.109)
Advancement	1.046*** (0.091)	1.059*** (0.080)	1.057*** (0.279)	1.183*** (0.209)
Responses	2768	3824	400	672
Proportion	0.534	0.528	0.490	0.519

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's logit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 1: Example of Job Offer Pair

**Based on the information provided, which offer would you be most likely to accept?
Please choose by clicking one of the buttons below:**

(3 of 20)


Congratulations!

My name is Jasmin and I will be your new manager! I have included some more information about the job below.

Because of the importance of any decisions, I expect that you will need my approval about once a week.

If everything goes well, I could see you move up in about one year's time.

The position pays about \$19.00 per hour.



Jasmin Dixon


Congratulations!

My name is Jake and I will be your new manager! I have included some more information about the job below.

Because of the importance of any decisions, I expect that you will need my approval about once a week.


If everything goes well, I could see you move up in about two year's time.

The position pays about \$17.50 per hour.



Jake Fischer

← →

0%  100%

Notes: One of the 20 choices each respondent answered in their browser (Internet Explorer displayed). Respondents indicated their preference by clicking one of the radial buttons at the bottom of each question.

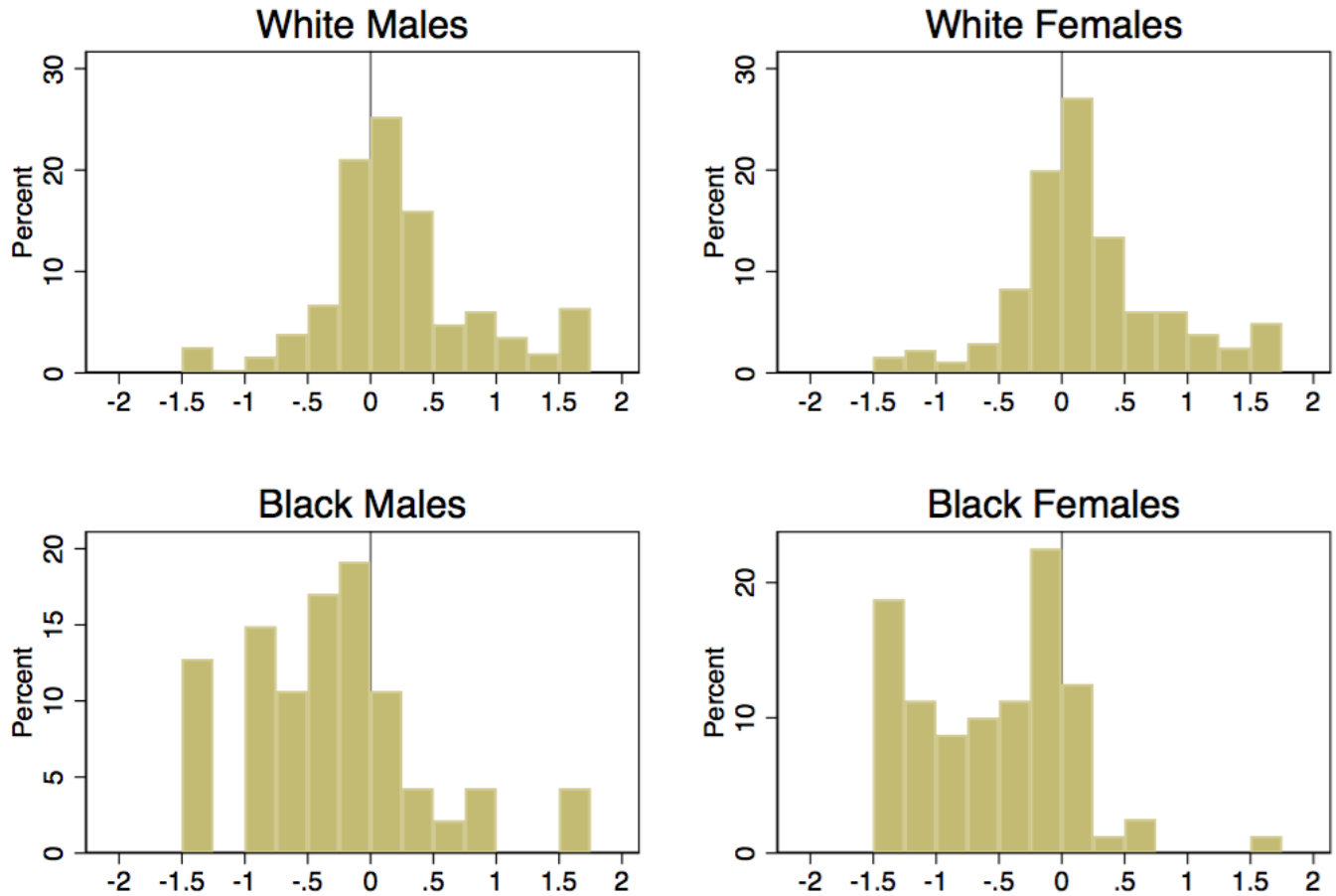
Figure 2: Photograph Pools



Notes: There are 80 manager photographs in total, with equal numbers over race and gender. The software draws one of the faces randomly from the pool required by a question. (If the respondent was meant to see a black male manager, then one of the 20 black male photos would be drawn.)

Figure 3: Individual WTPs - White Offer

WTP for White Manager



Notes: Histogram of Individual WTP for a white manager by respondent race and gender. Estimated by multinomial logit on all 20 questions per respondent. Multinomial probit estimations have similar moments (in appendix). Inclusion of the fixed tasks reduces the variance in the estimates. We present the lowest variance specification here. Bins are 0.25 USD wide. WTPs winsorized at the 5% level.

2.6 Appendices

Attribute	Level
Manager Race	White photograph and name. Black photograph and name.
Manager Gender	Male photograph and name. Female photograph and name.
Independence	Need approval once per week. Need approval twice per week.
Advancement	Promotion in about 1 year. Promotion in about 2 years.
Income	\$20.50 per hour. \$20.00 per hour. \$19.50 per hour. \$19.00 per hour. \$18.50 per hour. \$18.00 per hour. \$17.50 per hour.

Table A1: Respondent Summary Statistics

		White		Black		Whole Sample
		Male	Female	Male	Female	
Age	Average	35.65	37.01	35.22	33.23	36.09
Income	\$0 - \$24,999	22.22	20.31	17.65	14.81	20.35
	\$25,000 - \$49,999	30.86	34.66	29.41	29.63	32.56
	\$50,000 - \$74,999	22.53	21.41	27.45	35.80	23.43
	\$75,000 - \$99,999	10.19	12.36	9.80	8.64	11.11
	\$100,000 or more	14.20	11.26	15.69	11.11	12.54
Marital Status	Single	50.62	37.31	50.98	48.15	43.78
	Married	43.52	49.45	41.18	41.97	46.20
	Separated	5.86	13.25	7.84	9.88	10.01
Education	Highschool or Less	12.96	10.60	13.72	7.41	11.33
	Some College	24.69	24.50	37.25	28.40	25.63
	Associate's Degree	8.02	16.34	3.92	20.99	13.09
	Bachelor's Degree	38.89	32.67	27.45	30.86	34.43
	Graduate Degree	15.43	15.89	17.64	12.35	15.51
Labor	Full Time	61.11	43.93	56.86	48.15	51.16
	Part Time	11.42	18.76	9.80	11.11	14.96
	Contract	3.09	3.75	3.92	6.17	3.74
	Student	10.50	9.94	15.69	12.34	10.67
	Laid Off	0.62	0.44	1.96	0.00	0.55
	Looking	7.10	13.02	9.80	9.88	10.45
	Out of Labor Force	6.17	10.15	1.96	12.35	8.47
Highest Priority	Income	76.85	79.25	80.39	79.01	78.44
	Independence	4.32	3.75	3.92	3.70	3.96
	Advancement	16.67	15.45	13.73	17.28	15.95
	Manager	2.16	1.55	1.96	0.00	1.65
Lowest Priority	Income	2.16	1.32	3.92	1.23	1.76
	Independence	35.80	30.91	25.49	38.27	33
	Advancement	7.41	6.62	19.61	6.17	7.59
	Manager	54.63	61.15	50.98	54.32	57.64
Hispanic	Yes	5.86	7.28	5.88	6.17	6.6
	No	94.13	92.71	94.12	93.83	93.40
Applied to job? (Last 12 months)	Yes	53.40	52.76	49.02	51.85	52.7
	No	46.60	47.24	50.98	48.15	47.3
Will apply to job? (Next 12 months)	Very Likely	33.02	33.77	33.33	29.63	33.11
	Likely	25.62	26.49	39.22	32.10	27.39
	Not Likely	41.36	39.74	27.45	38.27	39.49
Respondents		324	453	51	81	909

¹ Married or common law. ² Separated, divorced or married. Black males and white males are not statistically different in household income while white females and black females are.

Table A2: White and Black Manager Photo Comparison

White Coefficient	Both	Males	Females
Age	-0.502 (1.155)	0.558 (1.848)	-1.491 (1.441)
Afraid	-0.053 (0.078)	-0.120 (0.103)	0.013 (0.117)
Angry	0.074 (0.123)	0.206 (0.177)	-0.049 (0.172)
Attractive	0.087 (0.131)	0.039 (0.188)	0.133 (0.187)
Babyface	-0.155 (0.125)	-0.202 (0.186)	-0.115 (0.171)
Disgusted	0.019 (0.093)	0.123 (0.128)	-0.076 (0.135)
Dominant	-0.087 (0.128)	0.098 (0.156)	-0.282 (0.179)
Feminine	-0.001 (0.282)	0.032 (0.091)	0.087 (0.155)
Happy	-0.005 (0.119)	-0.036 (0.162)	0.025 (0.178)
Masculine	0.005 (0.275)	-0.137 (0.124)	0.023 (0.128)
Prototypic	-0.163 (0.162)	0.016 (0.253)	-0.324 (0.204)
Sad	-0.103 (0.109)	-0.194 (0.146)	-0.011 (0.162)
Suitability	0.001 (0.126)	0.209 (0.203)	-0.196 (0.151)
Surprised	-0.055 (0.053)	-0.067 (0.081)	-0.042 (0.071)
Threatening	0.031 (0.100)	0.187 (0.138)	-0.128 (0.132)
Trustworthy	-0.031 (0.070)	-0.130 (0.106)	0.069 (0.090)
Unusual	-0.048 (0.094)	-0.138 (0.131)	0.031 (0.134)
T^2 p-value	0.593	0.180	0.236
Photos	80	40	40

Presented are the coefficients of white photographs on each of the subjective characteristics. Characteristics are on a 7 point scale, with the exception of age. For example, the 40 white photograph subjects are on average 0.50 years younger than their black counterparts. White male subjects are 0.55 years older than black male subjects.

Table A3: Male and Female Manager Photo Comparison

Male Coefficient	Both	White	Black
Age	-0.315 (1.156)	0.695 (1.505)	-1.354 (1.790)
Afraid	-0.057 (0.078)	-0.120 (0.100)	0.013 (0.121)
Angry	-0.036 (0.123)	0.086 (0.172)	-0.169 (0.177)
Attractive	-0.002 (0.131)	-0.050 (0.209)	0.044 (0.160)
Babyface	0.075 (0.126)	0.037 (0.177)	0.124 (0.181)
Disgusted	-0.064 (0.093)	0.032 (0.130)	-0.166 (0.135)
Dominant	0.394*** (0.120)	0.581*** (0.172)	0.202 (0.165)
Happy	0.001 (0.119)	-0.028 (0.154)	0.032 (0.187)
Prototypic	-0.170 (0.162)	0.000 (0.255)	-0.340* (0.194)
Sad	-0.122 (0.109)	-0.208 (0.138)	-0.026 (0.171)
Suitability	0.009 (0.126)	0.206 (0.176)	-0.199 (0.180)
Surprised	-0.022 (0.053)	-0.033 (0.063)	-0.009 (0.088)
Threatening	0.239*** (0.096)	0.391*** (0.141)	0.076 (0.128)
Trustworthy	-0.115* (0.069)	-0.211** (0.093)	-0.012 (0.102)
Unusual	0.106 (0.093)	0.025 (0.131)	0.194 (0.135)
Black Proportion	-0.073 (0.107)	-0.002* (0.001)	-0.101 (0.065)
White Proportion	0.045 (0.105)	0.041 (0.065)	0.003 (0.002)
T^2 p-value	0.000	0.002	0.037
Photos	80	40	40

Presented are the coefficients of male photographs on each of the subjective characteristics. Characteristics are on a 7 point scale, with the exception of age. For example, the 40 male photograph subjects are on average 0.315 years younger than their female counterparts. White male subjects are 0.695 years older than white female subjects.

Table A4: Fixed Task 2 - Attention Check

	1	2	3	4	5
White	0.000 (0.014)		0.020 (0.018)	0.019 (0.017)	0.021 (0.017)
Male		0.015 (0.010)	0.059** (0.027)	0.046* (0.026)	0.049* (0.026)
White * Male			-0.051* (0.029)	-0.041 (0.028)	-0.047* (0.028)
Constant	0.023* (0.013)	0.017*** (0.007)	-0.000 (0.017)	0.249*** (0.041)	0.204*** (0.065)
Respondents	909	909	909	909	909
Job Priority Controls				YES	YES
Demographic Controls					YES

Notes: Depicted are coefficients of a linear probability model with dependent variable indicating a respondent's failure of the attention check. Fixed task 2 presents two jobs that are identical with the exception of wage. A failure occurs if the respondent selects a wage of \$17.50 in place of \$20.50 per hour. Respondents that indicated they are both white and black removed. Respondents stated their job attribute preferences at the end of the experiment. Demographic controls include marital status, education and employment. Standard errors in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table A5: Type Probabilities - Multinomial Probit

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.096*** (0.015)	0.064*** (0.013)	-0.082** (0.040)	-0.147*** (0.037)
Male Manager	-0.006 (0.012)	-0.026** (0.012)	-0.021 (0.027)	-0.012 (0.029)
Independence	0.049*** (0.016)	0.108*** (0.012)	0.070** (0.030)	0.112*** (0.026)
Advancement	0.310*** (0.023)	0.384*** (0.022)	0.200*** (0.042)	0.408*** (0.051)
Income	0.155*** (0.009)	0.180*** (0.00*)	0.120*** (0.018)	0.186*** (0.016)
Respondents	324	453	51	81
Responses	5184	7248	816	1296

Data are drawn from the 16 random questions. Coefficients are predicted changes in acceptance probability of a job offer. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Income is per 0.50 USD increase. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A6: Type WTPs - Multinomial Probit

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.310*** (0.050)	0.179*** (0.036)	-0.340** (0.148)	-0.396*** (0.095)
Male Manager	-0.020 (0.039)	-0.071** (0.032)	-0.088 (0.117)	-0.034 (0.078)
Independence	0.158*** (0.049)	0.299*** (0.034)	0.290*** (0.127)	0.300*** (0.070)
Advancement	0.996*** (0.065)	1.066*** (0.052)	0.835*** (0.178)	1.097*** (0.135)
Respondents	324	453	51	81
Responses	5184	7248	816	1296

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's probit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A7: Type WTPs - Mixed Multinomial Logit

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.312*** (0.047)	0.187*** (0.034)	-0.378*** (0.154)	-0.394*** (0.091)
Male Manager	-0.031 (0.037)	-0.072*** (0.031)	-0.110 (0.115)	-0.036 (0.073)
Independence	0.164*** (0.045)	0.287*** (0.032)	0.334*** (0.136)	0.303*** (0.066)
Advancement	0.987*** (0.064)	1.050*** (0.052)	0.807*** (0.185)	1.076*** (0.137)
Respondents	324	453	51	81
Responses	5184	7248	816	1296

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's mixed logit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The Mixed Multinomial Logit (McFadden and Train 2000) is also known as the Random-Parameters Logit Model (Cameron and Trivedi 2005)

Table A8: Type WTPs - Attention Check Failures Removed

	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.289*** (0.045)	0.189*** (0.033)	-0.410*** (0.133)	-0.394*** (0.091)
Male Manager	-0.028 (0.035)	-0.074*** (0.030)	-0.150 (0.111)	-0.036 (0.073)
Independence	0.193*** (0.042)	0.279*** (0.031)	0.280** (0.123)	0.303*** (0.066)
Advancement	0.980*** (0.061)	1.039*** (0.050)	0.823*** (0.181)	1.076*** (0.136)
Respondents	5040	7104	768	1296
Proportion	0.972	0.980	0.941	1.000

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's logit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A9: Type WTPs - Fatigue

Tasks 1 - 8	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.268*** (0.056)	0.142*** (0.046)	-0.603*** (0.235)	-0.441*** (0.116)
Male Manager	-0.088 (0.057)	-0.022 (0.043)	-0.044 (0.198)	-0.030 (0.093)
Independence	0.134** (0.065)	0.290*** (0.045)	0.314 (0.210)	0.274*** (0.092)
Advancement	1.080*** (0.082)	1.072*** (0.067)	1.089*** (0.274)	1.075*** (0.182)
Responses	2592	3624	408	648
Proportion	0.500	0.500	0.500	0.500

Tasks 9 - 16	Respondent Type			
	White		Black	
	Male	Female	Male	Female
White Manager	0.348*** (0.060)	0.235*** (0.042)	-0.127 (0.167)	-0.344*** (0.101)
Male Manager	0.019 (0.047)	-0.119*** (0.039)	-0.135 (0.143)	-0.047 (0.099)
Independence	0.193*** (0.051)	0.288*** (0.040)	0.313*** (0.134)	0.333*** (0.086)
Advancement	0.901*** (0.071)	1.028*** (0.057)	0.613*** (0.220)	1.081*** (0.159)
Responses	2592	3624	408	648
Proportion	0.500	0.500	0.500	0.500

Data are drawn from the 16 random questions. Estimates are generated by dividing that attribute's logit model coefficient by two times the income coefficient. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

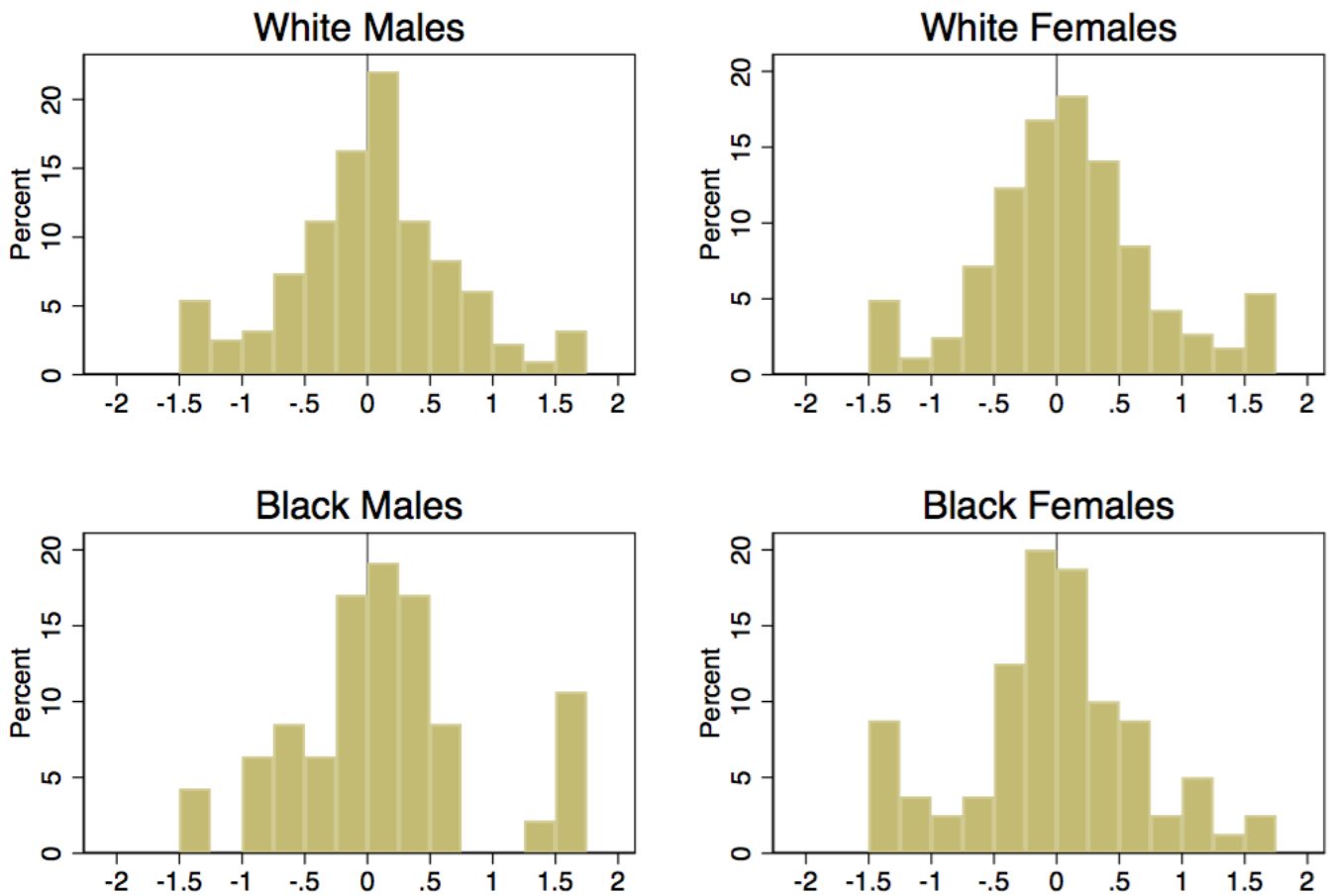
Table A10: Photo Popularity

	(1)	(2)	(3)	(4)
	White Males	White Females	Black Males	Black Females
White Manager	0.331*** (0.075)	0.171*** (0.066)	-0.485** (0.200)	-0.896*** (0.193)
Male Manager	-0.008 (0.054)	-0.082 (0.053)	-0.066 (0.113)	-0.010 (0.127)
Independence	0.242*** (0.068)	0.481*** (0.055)	0.312** (0.131)	0.536*** (0.120)
Advancement	1.435*** (0.109)	1.782*** (0.100)	0.880*** (0.193)	1.891*** (0.235)
Income	0.725*** (0.046)	0.845*** (0.041)	0.527*** (0.090)	0.881*** (0.084)
Photo Popularity	Y	Y	Y	Y
Responses	5184	7248	816	1296
Respondents	324	453	51	81

Data are drawn from the 16 random questions. White manager is relative to black manager. Male manager is relative to female manager. Independence is once a week relative to twice. Advancement is one year relative to two years for a promotion. Standard errors clustered at the respondent level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure A1: Individual WTPs - Male Offer

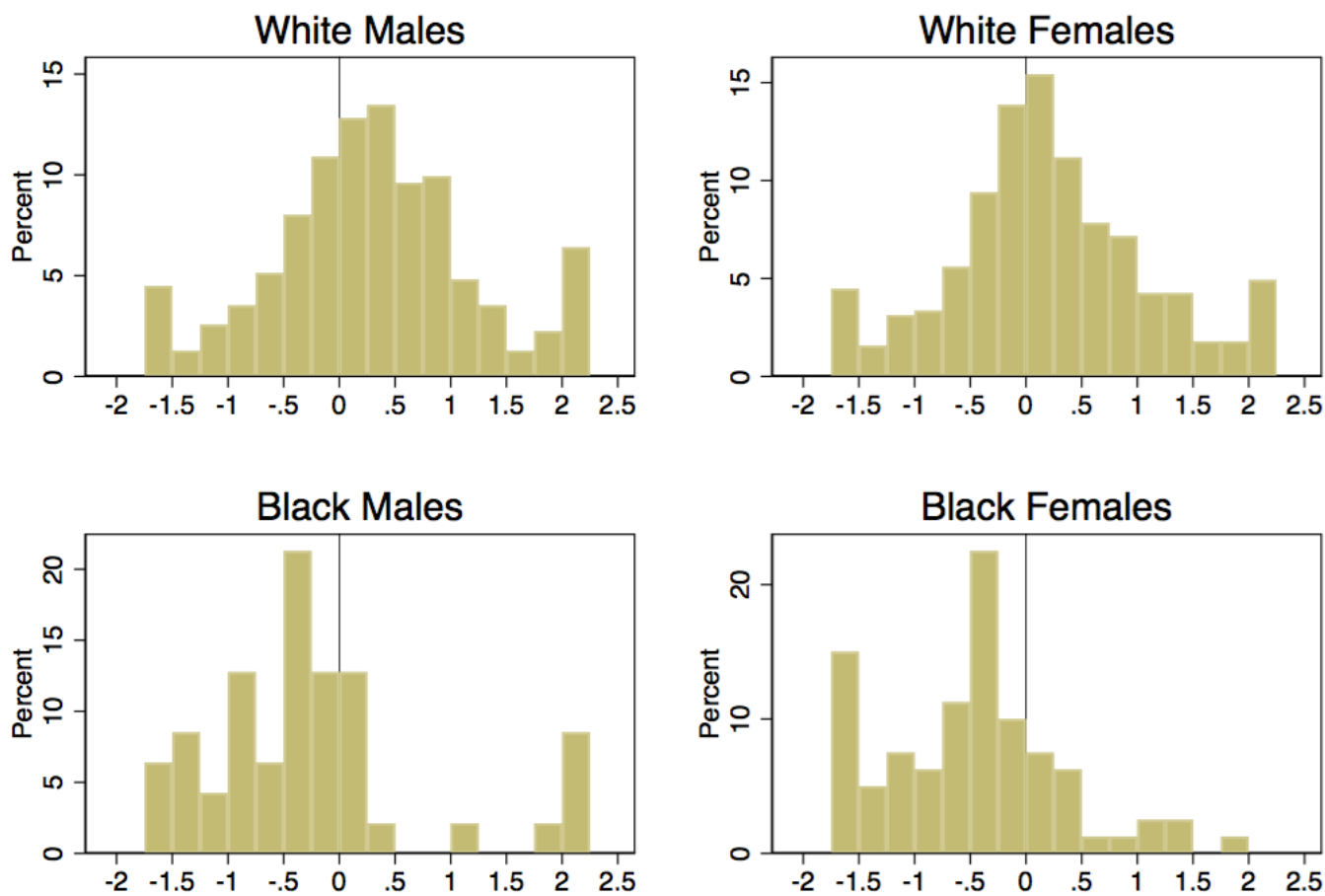
WTP for Male Manager



Notes: Histogram of Individual WTP for a male manager by respondent race and gender. Estimated by multinomial logit on all 20 questions per respondent. Multinomial probit estimations have similar moments. Inclusion of the fixed tasks reduces the variance in the estimates. We present the lowest variance specification here. Bins are 0.25 USD wide. WTPs winsorized at the 5% level.

Figure A2: Individual WTPs - White Offer - Logit 16 Questions

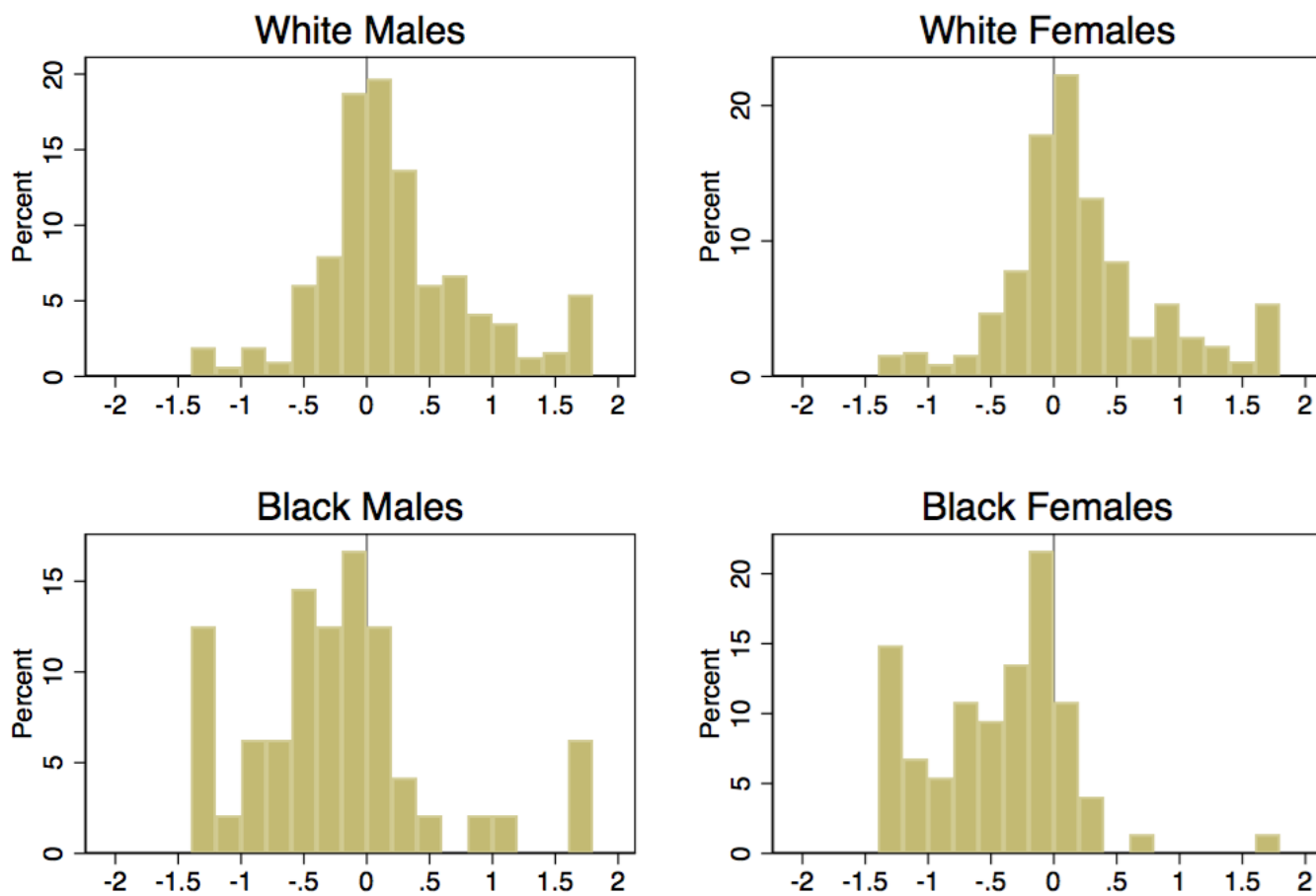
WTP for White Manager



Notes: Histogram of Individual WTP for a white manager by respondent race and gender. Estimated by multinomial logit on 16 idiosyncratic questions per respondent. Bins are 0.25 USD wide. WTPs winsorized at the 5% level.

Figure A3: Individual WTPs - White Offer - Probit

WTP for White Manager



Notes: Histogram of Individual WTP for a white manager by respondent race and gender. Estimated by multinomial probit on all 20 questions per respondent. Bins are 0.25 USD wide. WTPs winsorized at the 5% level.

Chapter 3

Student Aid Increases Performance and Decreases Graduation: Evidence from the 30% Off Ontario Tuition Grant

3.0.1 Abstract

A surprise 30% reduction in tuition increased academic performance for already enrolled post-secondary students – particularly those in STEM. I apply a difference-in-differences identification strategy at the Ontario-Quebec border combined with student fixed effects. The setting provides a uniquely plausible control group of *local* out-of-province students. Using student fixed effects leverages the eligibility of cohorts already in study, removing identification challenges that (likely unobservable) changes to the student body could pose in other settings. When treated, students perform approximately 0.09 standard deviations better than their peers. They are also less likely to exhibit unsubsidized behaviors such as part-time or summer study. Students who chose to enroll *after* the Grant was announced had lower graduation rates and lower admission averages despite no appreciable differences in income or original distance from campus, consistent with an increase in *access* to post-secondary education.

3.0.2 Thanks

I am grateful to Abel Brodeur, Anthony Heyes, Adam Lavecchia, Louis-Philippe Morin, and Matt Webb for helpful feedback. Errors are mine.

3.0.3 Ethics and Collaboration

This research was completed with administrative data accessed under University of Ottawa Research Ethics Board file number 11-17-15.

3.1 Introduction

Does student performance respond to no strings attached aid?

This chapter presents evidence that student aid increases performance of post-secondary students in a setting with high access and modest tuition. It is the first, to the best of my knowledge, to use comparable *local* students (subject to a different province's aid scheme) to identify the effects of student aid in any context. It is also the first, again to the best of my knowledge, to causally evaluate the \$500 million per year Ontario Tuition Grant. The Grant was implemented during the 2011 academic year with both new and continuing post-secondary students eligible for up to 30% off their tuition.¹ The Grant created cohorts of Ontario students with a subsidized education who had enrolled at a higher net price. Combined with administrative data from a setting in which a substantial number of local students are subject to another province's unchanged student aid scheme, I can cleanly identify the effects of student aid, using a difference-in-differences design with student (panel) fixed effects.

I find a modest increase in financial aid positively and significantly affects student outcomes within-student (how does a particular student fare after having their tuition reduced?) absent the often contaminating effects of between-student changes (how does a lower net price affect the type of student enrolled?). I find that treated students increase their performance through higher course grades, less failed courses, and reduced time on academic probation. I also find students exhibit less unsubsidized behaviors: both part-time and summer study for existing students fall in response to the Grant's restrictions.²

I also estimate that students who receive the Grant are more likely to enrol in courses with others who receive it (and those who do not are more likely to enrol in courses with others who do not). These concentration effects seem to benefit students – as the concentration of Ontario students increases, other Ontario students' performance is increased.

Later cohorts of students, who face a lower net price for education during the enrolment decision, had lower graduation rates, lower admission averages and were less likely to persist

¹In most of Canada, a distinction is made between colleges (which typically offer two year programs of an applied nature) and universities (which typically offer four year programs at the undergraduate level).

²Using administrative data for this study enables me to measure interesting variables that are often not available in other settings Figlio et al. [2016].

into their second year.³

The University of Ottawa (a large research, comprehensive, research intensive university on the Ontario side of the Ontario-Quebec border) provides a unique opportunity to study the effects of a grant with broad eligibility as it sources a sizable portion of its local students from the adjacent province of Quebec.⁴ Using administrative data (student age, six digit postal code, immigration and citizenship status) I construct both treatment and control group in order to use a difference-in-differences identification strategy. This out-of-province control group likely resembles the treatment group more naturally than other settings.⁵

The remainder of this chapter is structured as follows. In Section 2, I provide some context on student aid circa 2011, detail the Grant's eligibility, and provide a brief review of relevant literature. In Section 3, I describe the data used. In Section 4, I discuss the identification strategy and econometric models. In Section 5, I present results. Section 6 concludes.

3.2 Context, the Grant, and Literature Review

In this section, I first provide context of the student aid landscape when the Grant was introduced. Second, I detail the Grant and its eligibility. Lastly, I provide a brief review of recent and relevant literature.

3.2.1 Ontario Student Aid Circa 2011

In 2011, Ontario's 63% post-secondary education rate was one of the highest in the world and seven out of ten jobs were expected to need some form of post-secondary education.⁶ In

³In contrast to Denning [2017] who finds the marginal student is as likely to graduate under tuition reductions of Texas community colleges.

⁴For example, it is a 1.7 km drive from the University of Ottawa main campus to the Ontario-Quebec border via the Alexandria Bridge. Additionally, in a 50 km radius the median Quebec student's address is 9 km away from campus whereas the median Ontario student is 12 km away.

⁵Knight and Schiff [2019] formalize this concept with a model of out-of-state students not attending the school that is best for them due to tuition distortions (which are not present in my context). Further, even the labor market in Ottawa-Gatineau is dominated by a single employer (the Government of Canada employs 20% of the Ottawa-Gatineau workforce), reducing differences between the treatment and control group's returns to education that could be caused by employer heterogeneity in other settings, a channel recently documented in Engbom and Moser [2017].

⁶Ontario Government 2011

the 2008 academic year, 360,000 students were working on undergraduate degrees full time in Ontario. In 2011, 397,851 students were enrolled. By the 2014 academic year, 413,490 students were completing degrees in Ontario representing a linear growth rate of 2.5% per year since 2008.⁷ While demand for education was rising, available government grants and loans did not always cover costs.⁸

The provincial political landscape was particularly stable during this time (the level of government responsible for education administration in Canada). The Ontario Liberal party would be in power from 2003 through 2018 - 5 years prior to the beginning of my sampling period and after winning two consecutive elections.⁹ Education is one of the main campaign policy pillars for Ontario provincial parties; in the 2011 election (wherein the Grant was announced) the Liberals were re-elected with an official platform that read: “We’re going to support all middle-class Ontario families with a 30 per cent across-the-board post-secondary undergraduate tuition grant. That means—every year—the families of five out of six students will save \$1600 per student in university and \$730 per student in college.”

3.2.2 The 30% Off Ontario Tuition Grant

In September 2011, the provincial government announced the 30% Off Ontario Tuition Grant, which would reduce by 30% the average tuition paid by an Ontarian student at an Ontario Institution. At its beginning, the Grant offered \$800 per semester to eligible university and college degree students, up to a maximum of \$1,600 per year (rising regularly to \$1900 in its final year of 2016-2017). During the 2011-2012 academic year, University of Ottawa tuition and fees for a full time undergraduate student in Arts were just over \$6000, meaning the Grant produced an effective reduction of around 27%. For an engineering student paying \$7,800 the Grant relieved only 20%. The Grant was funded via “efficiencies and savings” rather than through a redistribution of student aid and ultimately discontinued in 2017. After its announcement, the Grant would be effective only three months later, with both new (beginning in Fall 2012) and current students eligible as long as:

- they were a full-time student at a public college or university in Ontario,
- they were in a first entry program,

⁷Statistics Canada Table 3710001101.

⁸Through the primary source of student funding, the Ontario Student Assistance Program, students could expect to receive a maximum of \$150 per week of study (full time students, single, and with no dependents) in loans. In 2010, tuition at the University of Ottawa was around \$6,480, well above the loans maximum. For the 2011 academic year, once the OTG was announced, tuition fell to \$5,730.

⁹Elections in Ontario were held in 2003, 2007, 2011, 2014, and 2018.

- their parents' gross income was \$160,000 or less,¹⁰
- they were a Canadian citizen, a permanent resident or a protected person,
- they were an Ontario resident, and,
- they had left high school within the last four years.¹¹

A total of 310,000 students would receive the Ontario Tuition Grant in its first year.¹²

3.2.3 Student Aid In Other Contexts

Researching whether student aid 'works' is recently a case of which program and which outcome is measured [Dynarski and Scott-Clayton, 2013]¹³ The majority of student aid research evaluates policies on its effects on student performance, persistence, and completion.

Performance, how well students are doing in their courses, is often the most detailed measure of student success available. While the function connecting human capital accumulation (learning) and course grades is not necessarily one-to-one, it allows researchers to determine if a program is having at least some measurable effect on student success. For example, Angrist et al. [2009] used an experiment to improve academic performance of first year university students at the University of Toronto. Treatments consisted of academic support services, financial incentives, or both. They found that only the combined treatment had an effect (and only for women) of higher grades and less time spent on academic probation. Leach et al. [2010] study Ontario standardized exam grades and finds effects based on funding re-evaluations from consolidations of school boards. Notably, wealthier school boards' scores are harmed while relatively poorer school boards' scores benefit from school board consolidation. Funding is also studied by Card et al. [2010] who look at Ontario primary school test scores and competition between publicly funded sectarian and Catholic schools, finding that competition could increase scores by 6-8% of a standard deviation.

Persistence and completion are focal outcomes of student aid research as they are directly rewarded in the labor market. Research on student aid and completion typically finds that aid encourages completion. For example, Dynarski [2003] uses an unexpected policy change to identify the effects of a reduction of aid (the elimination of the American Social

¹⁰Over 80% of students in Ontario had parents making less than \$160,000. Link

¹¹Students who graduated high school before January 2008 are never eligible. In 2014, the OTG was extended to fifth-year co-op students, who had previously been ineligible for funding after four years of their program.

¹²There was no additional application required for the grant - any student who had applied for a government load was automatically considered for the Grant - likely adding to its substantial take-up [Dynarski and Wiederspan, 2012].

¹³See also for a thorough overview of lessons learned from student aid implementations. See Page and Scott-Clayton [2016] for additional policy prescriptions.

Security Student Benefit Program in 1982). She finds that removing previously generous assistance to students corresponds to a reduction in college attendance and completion. More recently, Carlson et al. [2019] use a pre-registered study using randomization and finds that student aid increased persistence by 1.7 percentage points for 4 year university students, with little variation of treatment effect attributable to cohort, race, gender or receipt of food stamps. Studying relatively large increases in student aid to U.S. veterans following the Post-9/11 GI Bill (and leveraging the announcement's timing) Barr [2019] finds large degree attainment increases from student aid even for individuals with already high levels of support. Castleman and Long [2016] study the Florida Student Access Grant, a need based grant with sharp eligibility criteria. They find that student aid increased credit accumulation and increased graduation rates by 22% for students near the cutoff. As in my setting, Bettinger et al. [2019] examines student aid wherein eligibility was not known ex-ante. They find that increasing aid encourages both undergraduate and graduate degree completion. Finally, Denning [2019] studies the effects of student aid on Texas university students and finds aid increases credits attempted and reduces in-school earnings, accelerating time-to-degree.¹⁴

3.3 Data and Sample Restrictions

I begin with grades for over 2.6 million completed courses. The sample includes any student who began studies after Fall 2007 until Fall 2019. I connect these grades to institutionally-provided student information such as age and six-digit postal code (to determine Grant eligibility), or sex and program of study (to examine heterogeneity). Data on financial status of a student comes from the 2016 Canadian Census of Population.

Because my goal is to identify effects of the Grant, I need to restrict my sample to those who actually experienced it and a plausible control group. I keep only students from Ontario (74% of population) or Quebec (15% of population). As the Grant was only payable to students who had been out of high school for less than four years, I sample only courses completed during the 2008-2014 academic years.¹⁵ I also restrict the analysis to students aged 22 years and below due to this restriction, as Ontario high school students normally graduate in the year they turn 18. As only Canadian citizens, permanent residents, or protected persons were eligible for the OTG, I also remove international students and

¹⁴In school earnings, unstudied here due to data limitations, would be an interesting empirical question - the Ontario Student Assistance Program restricted the amount of earnings students could earn using a \$1 for \$1 clawback rate.

¹⁵This helps mitigate a policy change introduced in 2014, when the Grant was extended to fifth-year co-op students who had previously been ineligible for funding after four years of their program.

students without Canadian citizenship from the analysis.

I make two additional sample restrictions. First, while my econometric models (difference-in-differences with panel fixed effects) would estimate an effect of the Grant only on treated group students who complete courses both in the pre and post periods (i.e. switchers), the control group to which they would be compared would not require the same criteria (making interpretation difficult at best). To maintain comparability between treatment and control, I restrict the sample to students who complete courses in both the pre and post periods. The post period begins in Winter 2012 - halfway through an academic year.

Second, while the Grant was only paid for semesters completed as a full time student, it is possible that a student (whom prior to the program's announcement was a part time student) may change their status to full time in order to benefit. In light of this potential and endogenous decision margin, I restrict the analysis to those who enrolled as full time students (reducing sample size by 1.45% of courses or 490 students).

With these restrictions in place, I use a sample of approximately half a million ($N = 590,039$) courses completed by 19,573 students.

3.3.1 Summary Statistics

Summary statistics relating to course performance and student characteristics are provided in Table 1. The average course grade is 74%, corresponding to a "B" in the university's official grading scheme (Table A2). Grades vary considerably; the overall standard deviation is 14%, or almost 3 letter grades. The within-student standard deviation (presented in square brackets) is 10%, or two letter grades around the mean.

Around 4% of courses taken result in a failing grade. Courses taken while under academic probation (a cumulative GPA below 64% - a "C") make up 21% of the sample. For students who enrolled as full time, 6% of studies are part time (4% if excluding summer, which make up 5% of the overall sample).

Most courses during this sample period are taken by students who graduate (note that there is no within-student variation of this and later variables). Admission averages are centered around a "B+". Students come from postal codes with average 2016 individual incomes (\$52,700) close to the 2016 average the Ottawa-Gatineau area (\$51,523), with large variation between students. Female students account for 62% of the data (higher than the Ontario proportion of 56%). STEM students account for 25% (in line with 25% of Ontario students studying STEM).¹⁶

Summary statistics by treatment group are presented in the following columns.¹⁷

¹⁶Statistics Canada Table 3710001101

¹⁷Because of the inclusion of summer study in this table I use 630,842 exams, rather than the smaller sample later used in estimation.

3.4 Identification and Econometric Models

Identification comes from the rapid deployment and retroactive eligibility of the Grant, which created cohorts of students who received it but had enrolled prior to its announcement.¹⁸ In other words, only students who would have attended university absent the Grant are examined and students who would attend only if given the Grant are not. By their absence, I can cleanly identify the effects of student aid without a potentially confounding effect of compositional changes found in other settings. Note that any estimates are derived from Grant (treatment) eligibility, and so are ostensibly estimates of intention-to-treat.

I use panel data and a difference-in-differences identification strategy to identify the effect of student aid on the treated (Ontario students) compared to a control group (Quebec students).¹⁹

For outcome variables such as course grade, which can vary *within* student, I use a student fixed effects model estimated by Ordinary Least Squares. This allows me to strip out the effects of any time-invariant student-level unobserved characteristics. Depending on the specification, I include a year trend, year fixed effects, neither, or both (much like in Besley and Burgess [2004]). The full specification is:

$$Outcome_{i,t} = \beta_0 + \beta_1 * (Treat \times Post) + \beta_2 Post + \gamma_i + T + \eta_t + \epsilon_{i,t} \quad (3.1)$$

Where $Outcome_{i,t}$ is (for example) the course grade for individual i completing a course in year t . My parameter of interest is β_1 , the difference-in-differences estimate, which is the coefficient of the treated group when treated (Ontario students post Winter 2012) compared to the control in the same period. Student fixed effects in combination with the DID design (and additional sample restrictions made) mean β_1 is estimated only on those who transitioned *from* untreated *to* treated. β_2 estimates the average change for both treated and control groups after Winter 2012.²⁰

The inclusion of a year trend removes any linear pattern common to both the treated and control groups, including, for example, grade inflation. Year fixed effects capture (possibly non-linear) changes between years that are common across groups. Standard errors are clustered at the student level.

For outcome variables such as graduation, which vary only *between* student, I rely more heavily on the common trend assumption of the DID design. In these cases, I estimate the

¹⁸Ontario students typically apply to university in January or February of the year they wish to begin fall classes. The September 2011 cohort had the second half of their first year subsidized, after making the decision to enrol without knowledge of the Grant. The first students to enrol after the announcement began studies in Fall 2012.

¹⁹It is important to note that there were no substantive policy changes to Quebec provincial aid during this time period [Ford et al., 2019].

²⁰Following the advice of Bertrand et al. [2004], I collapse pre and post periods.

following model:

$$Outcome_i = \beta_0 + \beta_1 * (Treat \times Post) + \beta_2 Post + T + \eta_t + \epsilon_i \quad (3.2)$$

Where $Outcome_i$ is (for example) an indicator variable which takes a value of 1 if the student graduated. β_1 will identify the difference between treated group students who enrolled *after* the program was enacted to treated group students who were enrolled before the announcement. β_2 is the average graduation rate for students (control and treated) who enrolled after Winter 2012. As there is only one time observation per student (the enrollment year), the year trend captures linear trends in *cohorts*, while year fixed effects capture cohort-specific non-linear shocks.

3.5 Results

First, I begin by analyzing measures of student performance: course grades, failed courses, and courses completed while subject to academic probation. While course grades and learning are important in their own right, they have also long been connected to later wages earned [Jones and Jackson, 1990]. I then examine heterogeneous effects on these performance measures. Students, regardless of outcome measure, perform better when given the Grant.

Additionally, I present evidence that the Grant changed the composition of classrooms. Within-student, treatment and control students became less integrated, ultimately raising the performance of both groups.

Second, I present analysis on how the Grant reduced unsubsidized behaviors. Since the Grant only reduced tuition for full time study of two semesters per academic year (where three 13-week semesters are available in Canada), both part time and summer study were unsubsidized. I find that these behaviors were reduced in students otherwise eligible for the Grant.

Third, I analyze graduation rates and find that students who had enrolled after the program was announced (and therefore faced a lower cost of education) graduated at a significantly *lower* rate than students who were initially paying more to attend university. I also find that these students have much lower admission averages. Both results are consistent with results from human capital theory: as tuition becomes less prohibitive, weaker potential students rationally choose to enter into post-secondary studies to benefit from later increased wages. In the appendices, I present null estimates that the Grant had no measurable effect on average student enrollment income or distance from home to the university.

3.5.1 Graphical Evidence

In Figure A2, I plot the average course grade (standardized across all students and years) by treatment group over the sample period of 2008-2014. A vertical line is included for the 2011 academic year, when the Ontario Tuition Grant was introduced. Quebec (control) students do consistently better throughout (likely due to the fact that university participation rates in Quebec are relatively low, despite easy access to it Finnie and Mueller [2017]). For both the control (dashes) and treated group (long dashes) there is a general upward trend. To make the difference obvious, I have plotted it separately (connected squares with values attached). The difference between treatment and control groups begins at 0.26 of a standard deviation and remains relatively stable until the 2011 academic year, when the Grant is introduced. The difference then falls to 0.2 and remains thereabout.

In Figure 2, I provide a similar analysis (and graphically testing my results much in the spirit of Kearney and Levine [2015]). I plot interaction terms of treatment status \times academic year from a regression of standardized course grade on treatment status, year fixed effects, and their interactions. I find that course grades for the treated group are statistically different from the base year (2011) for the years 2008, 2009, and 2010 separately (and together $p < 0.004$). In contrast, 2012, 2013 and 2014 are not statistically different from 2011, either separately or together ($p < 0.36$). Whiskers indicate 95% confidence intervals.

3.5.2 Course Performance Increase With Aid

In Table 2, I estimate the effect of the Ontario Tuition Grant on academic achievement. The dependent variable is standardized course grade.²¹ Presented coefficients are changes in standard deviations. As detailed in the Data section, the analysis is restricted to courses completed by students who move from the pre to the post period. Said differently, only students who completed courses prior to and after the Winter semester in 2012 are included.

In the first column, the *Post* coefficient captures the average change for both treated and control groups as they move from the pre to post periods; both groups do better. The *Post* \times *Treated* coefficient identifies the *additional* change in grades the treated group exhibits as it moves from the pre to the post period i.e. when they are treated. When a student is given 30% off of their tuition, they are estimated to do around one tenth of a

²¹Standardized across all students and years. I show robustness to alternative standardizations in Table A8. Course grades are recorded as one of ten letters, corresponding to percentage intervals (e.g. a score of 86 in a course earns a student an ‘A’: a score in the interval 85-89%. For each letter grade, I assign the percentage corresponding to the mid-point of its interval (see Table A2). That percentage is then standardized). This reporting granularity is an additional source of measurement error, notably not correlated with the treatment variable or its assignment. While such measurement error does not bias OLS estimates, it does increase the associated standard errors.

standard deviation better - quite large in the education literature (for example, Card et al. [2010] find an effect size of 8% of a standard deviation from school competition.)

In the second column, I introduce a year trend to account for any linear change in grades (within student). When introduced, the effect of the post period is no longer positive, suggesting that a significant portion of the *Post* coefficient in the first specification was due to linear increases in grades over time, rather than due to pre-post differences (a likely result but not a certain one, given Figure A2 is at the aggregate rather than within-student level). The introduction of the trend reduces the estimate of the $Post \times Treated$ coefficient, to 9% of a standard deviation and it remains highly statistically significant.

In the third column, in place of a linear year trend, I introduce year fixed effects. This specification allows me to remove shocks that are common across all students within a year. The $Post \times Treated$ coefficient is roughly the same as in the second column.

In the fourth column, when the specification contains both a linear trend and time fixed effects, the estimates are not disturbed. In the final column I present my preferred specification, wherein I include fixed effects for the levels of a students' year of study (a categorical and ordinal variable ranging from 1 to 4).²²

In Table 3, I investigate the relationship further. In every specification (except the last) I reproduce the preferred specification from column 5 in Table 2. I now restrict the size of the analysis 'window' on either side of the treatment introduction date. In the first column I reproduce the estimate from above with the full sample. In the second column, I remove exams taken in the 2008-2009 and 2014-2015 academic years, the furthest dates on either side of the treatment's introduction. The treatment effect is substantively the same as in the whole sample, and the same occurs in the third column with a further restricted sample. In the fourth column, the sample is restricted to exams taken during the single 2011-2012 academic year (when the OTG was implemented). I can no longer include trend and year fixed effects because of the restriction to a single year. The $Post \times Treated$ estimate is not statistically different from zero, with a large reduction in effect size and a standard error of the same magnitude as the first three columns. While this estimate may be troubling without additional context - why should the program not be immediately effective - this was also the only academic year in which students received \$800 (rather than the \$1600

²²Included to address whether the results are driven by a possible quirk in the control group. Quebec students often attend university after three years (rather than four years for the treated group) of secondary school and after two years of CEGEP (Collège d'enseignement général et professionnel - an intermediary step whose resulting diploma is required for Quebec university admission) providing some but not all Quebec students with enough credits to begin in second year at the University of Ottawa. The difference-in-differences design will account for this relationship as long as it is stable, if the proportion of Quebec students entering second year as opposed to first changes during the study period this should be adequately addressed by the inclusion of year-of-study indicators.

the following year) and the only year in which the Grant was given *after* most universities required tuition payments be paid, suggesting a possible mechanism through which the grant affects grades.²³

3.5.3 Academic Probation Decreases With Aid

In Table 4, I change the dependent variable to whether the student, at the time of completing the course, had not achieved a cumulative grade point average high enough to avoid academic probation. Successive semesters on academic probation trigger mandatory withdrawal from study, effectively halting degree progression.²⁴ Over 20% of courses are completed by students on academic probation. My most conservative estimate finds that when treated, students reduced the likelihood of being on academic probation by 2.9 percentage points, a reduction of 13%.

3.5.4 Failures Decrease With Aid

In Table 5, I use the same specifications as in Table 2 while changing the dependent variable from standardized course grade to an indicator variable that takes a value of one if the student fails a course and takes a value of zero otherwise. This dichotomous variable is likely more important (albeit more granular) than standardized course grade, as the consequences of course failure include a lack of progress and increased time towards degree completion, at a minimum. For all columns, I include the mean of the dependent variable; 3% of courses taken are failed. When treated by the Grant, the failure rate is reduced by 1.0 to 0.7 percentage points, depending on specification. This corresponds to a 21-30% reduction in course failures.

3.5.5 STEM Students Benefit More From Aid

Following results such as Angrist et al. [2009], in Table 6 I explore whether there are heterogeneous effects of the Grant for the effects presented in Table 2 (course grades), Table 4 (time spent on probation), and Table 5 (course failures). In each column, I use the fully enriched (preferred) model found in the rightmost column in each of the previous tables.

In the first column of Table 6, I introduce an indicator variable for whether a student is female. This necessarily changes the interpretation of the baseline coefficients. *Post*

²³For example, Manoli and Turner [2018] find that cash-on-hand significantly determines the decision of United States high school seniors to attend college by using timing shifts in tax refunds of the Earned Income Tax Credit.

²⁴Lindo et al. [2010] study the effects of being placed on academic probation in more detail. They find that probation can discourage persistence for some students and increase the GPAs of students who continue studies.

is now the change in standardized course grades for male students moving from the pre to post periods. The $Post \times Treated$ coefficient is the change in *male* students moving from untreated to treated. Neither the $Post \times Female$ nor the triple interaction term $Post \times Treated \times Female$ coefficients are practically or statistically significant. This means that the effects of the Grant are estimated to be equal for both male and female students.

In the second column of Table 6, I introduce an indicator variable for whether a student enrolled as a STEM student (began studies as a student in the Faculty of Science or in the Faculty of Engineering). The baseline coefficients now correspond to all Non-STEM students (which includes business, social sciences, humanities and the arts). The $Post$ and $Post \times Treated$ coefficients remain statistically significant, as Non-STEM students benefit from the Grant's introduction. The triple interaction term, which estimates the *additional* benefit the Grant had on STEM students, finds that STEM students benefited 77% more (despite the Grant relieving a *lower* proportion of their overall tuition burden).

In the third and fourth columns, I repeat the same exercise with the probation dependent variable discussed in Table 4. Male and female students equally reduce the number of courses spent while on academic probation when treated with the Grant. STEM students benefit thrice as much as non-STEM students along this performance measure.

In the fifth and sixth columns, I use the course failure indicator described for Table 5. Male and female students equally reduce the number of courses they ultimately fail. Surprisingly however, STEM students and non-STEM students are not differently affected by the Grant. When combined with the results in columns 2 and 4, this suggests that the Grant's effects are concentrated on non-marginal STEM students.

3.5.6 Wealthier Students Perform Better

In Table 7, I median-split students based on the average income of their six-digit postal code in the 2016 Canadian Census. I then perform a regression like those in Table 2 with a slight modification. I now separately estimate $Post$ and $Post \times Treated$ for upper and lower medians, while estimating a common year trend, year fixed effects and year of study fixed effects. Students in the *upper* median of the wealth distribution have a greater increase in their performance than those in the lower median. While this difference is only marginally statistically significant ($0.08 < P < 0.11$, depending on specification), the difference is practically significant at an almost 50% increase in benefit for the 'wealthier' students. This is consistent with context of the relative amounts of aid available at the time - the Grant was directly advertised as targeting the unmet needs of middle class students.

3.5.7 Unsubsidized Behaviors Decrease With Aid

In Table 10, I use semester as the unit of observation. The dependent variable is an indicator that takes the value one if that semester is taken at a partial or reduced academic load. This margin of adjustment is notable because as students move from full-time study to part-time, they become *ineligible* for the Grant; I measure whether students engage more or less in *unsubsidized* behaviors. As always, the sample is restricted to students that enrolled as full-time. Semesters taken during the summer are not included in this analysis. For this sample, around 8.1% of semesters are taken as part-time studies. The $Post \times Treated$ coefficient indicates that when full time study is subsidized, part time study is reduced by 1.4 percentage points (a 17% reduction).

In Table 11, I return to using courses as the unit of observation. This analysis is the only one to include summer courses. Approximately 6% of courses are taken during the summer, with 48% taken in Fall, and the remaining 46% in Winter. This margin of adjustment is of note because the Grant subsidized only two full time semesters per year. As always, the sample is restricted to students that enrolled as full-time. I find that depending on specification, there is a slight decrease of around 0.5 percentage points (a reduction of 9%) when students are treated during the Fall and Winter semesters, noting the effect is absent in the preferred specification.

3.5.8 Student Aid Has A Larger Effect In Winter

In Table 12, I break down the effects of the Grant by season - Fall and Winter - separately. The motivation is that in the Fall, funds are more readily available to students through summer savings. While government funding is released in relatively similar installments for Fall and Winter, the amount released in September is often larger. At the same time, the fees charged by the university are only slightly higher in September (for example a 2011 entering student in social sciences owed \$ 3,114.11 for the fall semester and \$ 2,931.45 for the winter semester). Scholarships administered by the University are awarded in equal installments.

The introduction of the Ontario tuition Grant had different effects by semester. While the $Post \times Treated$ estimate (indicating the treatment effect on treated students in the Fall semester) remains nearly the same as in Table 2, the triple interaction term estimates that the benefit of the Grant is around 40% higher in the Winter, regardless of specification. While treated students seem to do much better during the winter term when given more aid, there is not an accompanying reduction in dropout (presented in Table A9), suggesting careful interpretation of this result.

3.5.9 Student Aid Decreases Graduation Rates

In this subsection, I probe the effects of the Ontario Tuition Grant for outcomes that necessarily do not vary within a student; the unit of observation is the student. The interpretation of the coefficients change and now better resemble that of standard difference-in-differences. The *Post* variable will capture the difference between students who enrolled in the pre period to students who enrolled in the post period. The $Post \times Treated$ coefficient will capture any *additional* difference that occurs in the treatment group. Specifically, after removing the effects of time (by removing Quebec students' changes over time) how did outcomes change for Ontario students? This is important as Ontario students shifted from being students who enrolled *prior* to their knowledge of being subsidized to those who enrolled *after knowing* they would be subsidized. In effect, I will be estimating changes to the external margin - how did the average student change, given the cost of education fell?

In Table 13, the dependent variable is whether a student graduated. The sample is restricted to students who began anytime between 2008 and 2014. I use a difference-in-differences design (now without panel fixed effects). The *Post* variable captures the change in graduation rates common to both Ontario and Quebec students. Compared to students enrolling between 2008-2011, students who enrol between 2012 and 2014 are 10% less likely to graduate. The $Post \times Treated$ coefficient estimates Ontario students' additional 6 percentage point reduction in graduation rates. While this result seems initially counter intuitive - when students are less burdened by tuition they are less likely to graduate - in Table 14 I will show that the *type* of student changes as well.

In the second column, I introduce year fixed effects which will remove cohort shocks common across the treated and control groups. In the third column I introduce, in place of fixed effects, a linear trend in cohort. In the fourth column, my preferred specification, I include both year fixed effects and trend controls. Regardless of specification, when students enrol in the post period they are 6 percentage points less likely to graduate, corresponding to a 9% reduction. An additional note is that the *proportion* of incoming students from treated and control groups is relatively stable before and after the policy change.²⁵

...And Admission Averages

In Table 14, I speak to a mechanism that could drive a reduction of graduation rates. In all columns, I find that the admission averages of treated group students are significantly *lower* when the Grant is known when students enroll. This coincides with rationally attending post-secondary education. When training is costly, the marginal worker is indifferent to purchasing education and commanding a higher wage or working. When the cost of this

²⁵In 2008, 15.5% of new students were from Quebec. This is followed by 14.9% in 2009, 16.9% in 2010, 15.9% in 2011, 16.6% in 2012, 16.0% in 2013 and 17.8% in 2014.

training is reduced (without a reduction of the perceived or actual benefits), then additional workers opt-into education until the marginal worker (who now benefits less than the original marginal worker) is indifferent.²⁶

...Without Changing Student Wealth Or Location

Other mechanisms, such as changes in student wealth or geographic sourcing of the student pool are not detected. In Table A3, the dependent variable is the 2016 Census average income of the six digit postal code a student reports as their *current* address. This is asking ‘when subsidized, do students move to wealthier areas?’ In Table A4 I examine the change in 2016 average income of the *enrolment* six digit postal code. I now ask ‘when subsidized, do students come from wealthier areas?’. The answer to both is no. In Table A5 I find no effect of the Grant on students changing how far away from campus they live. In Table A6 I find no changes in the average distance of the areas students are enrolling from. While we see reductions in graduation rates following the introduction of the Grant, this seems to be due not to changes in income or geographic origins of students and mainly due to a weakening of the marginal student.

3.5.10 No Effect On Persistence Into Second Year

In Table 15, the dependent variable is whether a student persisted - did they return to studies after first year? While I notably do not have data whether students left the university after first year *and* graduated elsewhere, the difference-in-differences assumption remains that the retention rate would change at least in the same manner for control and treatment groups over time. Almost 90% of students persist into their second year. In Table 15, the sample includes students who began anytime between 2008 and 2014, with one observation per student. Regardless of specification, students who enrolled under a reduced tuition load (lower net price) are not statistically less likely to persist.

3.5.11 Robustness to Alternative Clustering and Standardization Strategies

In Table A7 I report the results of using alternative standard errors. So far I have applied standard errors clustered at the student level, corresponding with the panel structure of my data. It is likely that observations within student are correlated (even after accounting for individual fixed effects). In the first column, I report standard errors that are unclustered

²⁶This is congruent with recent experimental work in Bleemer and Zafar [2018], who find that when prospective students are updated with the returns to university, intended attendance increases by 0.2 standard deviations.

and do not consider any level of heteroskedasticity. In the second column, I provide standard errors that are clustered at the student level (used throughout the remainder of the chapter). I find there is not a meaningful change in standard error size. In the third column, I cluster by the first term a student enrolled at the university (differentiating, for example, students enrolling in Fall 2010 and Winter 2011 despite enrolling in the same academic year) clustering at what could be considered treatment level (treatment in first year could be different than treatment in second year, and different if a student had started in Winter than in fall) and cohort determines this inter-year pattern. In the fourth column, I cluster at the student cohort proper. The challenge for these two exercises is the low number of cohorts (clusters) available, forcing me to bootstrap the standard errors. While the standard errors as measured in this manner are three to five times larger, my effect estimate is still statistically significant at a level well beyond 5%. In the fifth column, I cluster by cohort \times province, following the advice that clustering should always be done at the treatment (in my case, province) level. Results are sustained. Throughout this exercise, standard errors are the subject of the bootstrapping.

In Table A8 I challenge the robustness of my main result by re-estimating my preferred specification using alternative standardizations (means and standard deviations derived from different sets of ‘peers’). In the first column, I present what has been used throughout the chapter; standardization is done between all students and all years included in estimation. In the second column, standardization is *within* group as the treatment and control groups on average achieve different scores. The treatment effect’s estimated magnitude and significance are not substantively disturbed. In the third column, standardization is between all students, because while standardization enables comparison of effect sizes between studies, another benefit is it allows for a measurement of how effective a policy changes a student’s performance *relative* to their peers. Standardization so far has included only the treatment and control groups and has necessarily ignored the presence of others in the classroom (such as foreign students and those from other provinces). Estimates are once again little disturbed. Because the possibility exists that the act of standardization itself gives rise to my estimated effect sizes and its significance, in the fourth column I do not apply any standardization - leaving the dependent variable in its raw score out of 100. Regardless of measure, student grades are statistically significantly increased by the introduction of the Grant.

Another important source of robustness comes from my claim that the effects are coming from *local* students. In Table A1 I repeat the main analysis (presented in Table 2) while restricting the sample to students with addresses within 50km of campus. In the specification with a year trend, year fixed effects, and indicator variables for student-year-of-study, I find that the OTG increased course performance by 0.075 standard deviations (in Table 2 the estimate was 0.089 standard deviations).

3.6 Conclusion

This chapter was the first to causally study the student effects of the \$500,000,000 per year 30% Off Ontario Tuition Grant. Using a difference-in-differences design, the treatment group (Ontario students attending university for the first time) had significantly better outcomes when treated. The unique setting on the border between two provinces with independent aid schemes allowed for the construction of a control group that better resembles the treatment group than is often possible in other settings. Using the rapid deployment and retroactive eligibility of the Grant, I identified student cohorts that were treated with reduced tuition fees but made their enrollment decision prior to the reduction's announcement. These cohorts allow for identification of the effect of reduced tuition, separate from compositional changes.

A modest increase in government aid (without performance incentives or any “strings attached”) saw effect sizes typical for programs of much greater cost per student. Although aid was distributed equally, its effectiveness differed by recipient. While men and women equally benefited from the Grant, STEM students saw comparatively larger increases in their course performance and reductions in their likelihood to be placed on academic probation.

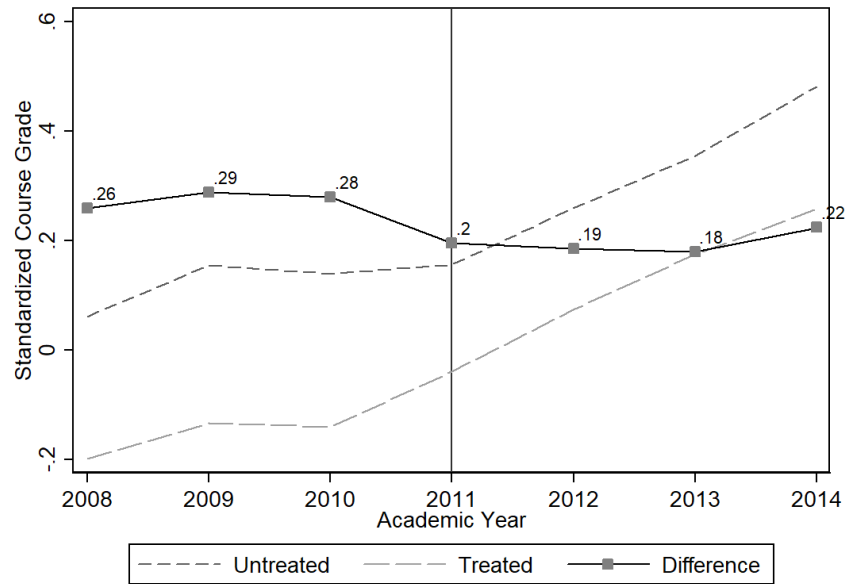
Students could endogenously become ineligible for the Grant if they reduced their course loads below full time or studied more than two semesters in an academic year; students who would otherwise be eligible were less likely to exhibit either behavior.

I find that weaker students, on average, attend studies when the sticker price is reduced. Comparing cohorts of students just before and just after the Grant was announced finds weaker students are less likely to graduate while having comparable family incomes and coming from the same places.

As part of their election platform the Ontario government pledged in helping more than 300,000 (in its first year) eligible students by reducing their tuition by 30%. Broad eligibility rules coupled with automatic applications meant many students qualified for the grant. Leveraging an opportunity at the Ontario border, I find evidence that this student aid program induced positive and significant changes to students already enrolled, and increased access to university for students who may not have otherwise attended.

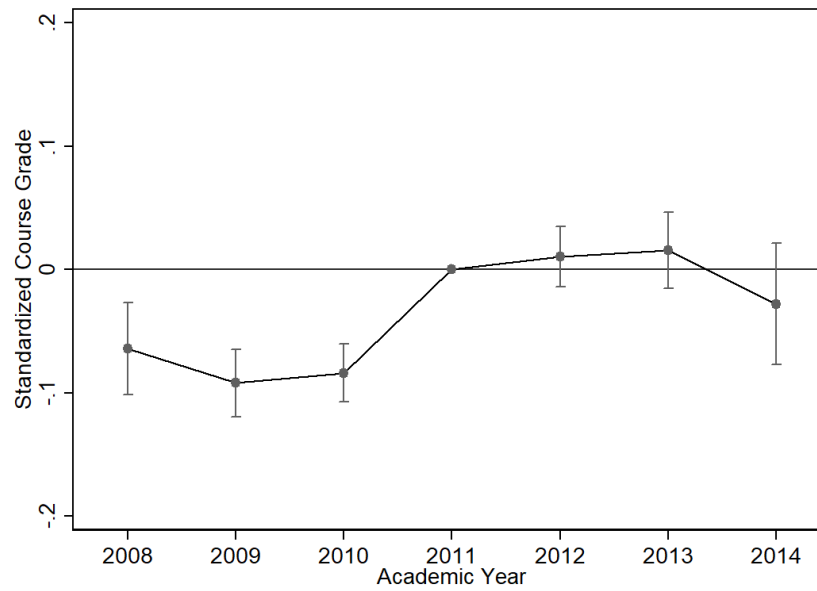
3.7 Tables and Figures

Figure 1: Trends by Treatment Group



Average standardized course grade by academic year. Treated students from Ontario. Control students from Quebec. Difference plotted as a solid line with values attached. Vertical line in 2011 provided when the Grant was announced and introduced. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Fall 2008 through Winter 2014.

Figure 2: Estimated Treatment Group Difference



Treatment \times *Year* coefficients from a regression of standardized course grades with treatment status, year fixed effects, and their interaction terms. The Grant was implemented in the 2011 academic year. 95% confidence intervals in whiskers. Treated students from Ontario. Control students from Quebec. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Fall 2008 through Winter 2014.

Table 1: Summary Statistics

	All	Treated	Control
Percent	73.92 (10.22) [10.46]	73.61 (10.20) [10.54]	76.50 (10.06) [9.76]
Fail	0.04 (0.11) [0.17]	0.04 (0.12) [0.18]	0.03 (0.10) [0.15]
Probation	0.21 (0.37) [0.22]	0.22 (0.38) [0.23]	0.13 (0.34) [0.18]
Part Time	0.06 (0.11) [0.22]	0.06 (0.11) [0.22]	0.06 (0.14) [0.22]
Summer	0.05 (0.07) [0.21]	0.05 (0.07) [0.21]	0.05 (0.08) [0.21]
Graduated	0.91 (0.37)	0.91 (0.37)	0.90 (0.38)
High Adm. Avg	0.75 (0.45)	0.77 (0.44)	0.60 (0.50)
Enrolment Avg. Income	52.70 (20.85)	52.88 (20.20)	51.14 (24.60)
Female	0.62 (0.487)	0.62 (0.49)	0.65 (0.48)
STEM	0.25 (0.42)	0.26 (0.43)	0.22 (0.39)
Observations	630842	563518	67324
Students	20075	17387	2688

Percent is the out of 100 score received in a course. Failed courses do not further degree progression. Academic probation is incurred if a student's grade point average is below a "C". Part-time study defined as three courses or less in a semester. Summer study runs from May-August. Graduated is defined as having graduated by 2019. Average income of the six-digit postal code from the 2016 Census. STEM students in the Faculty of Science or Faculty of Engineering. Students who move from the pre period to the post period. Domestic students aged 22 and below. Enrolled as full time only. Fall 2008 through Winter 2014. Between student standard deviation in parentheses, within-student standard deviation in square brackets.

Table 2: Effect of OTG on Course Grades (Standardized)

	(1)	(2)	(3)	(4)	(5)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score
Post=1	0.121*** (0.010)	-0.059*** (0.010)	-0.079*** (0.010)	-0.079*** (0.010)	-0.077*** (0.010)
Post=1 × Treated=1	0.134*** (0.011)	0.086*** (0.011)	0.085*** (0.011)	0.085*** (0.011)	0.089*** (0.011)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Exams	590039	590039	590039	590039	590039
Students	19573	19573	19573	19573	19573

The dependent variable is standardized course grade. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 3: Effect of OTG on Course Grades, Narrower Sample Windows

	(1)	(2)	(3)	(4)
	08-14	09-13	10-12	11-11
Post=1	-0.077*** (0.010)	-0.073*** (0.010)	-0.059*** (0.010)	-0.024** (0.012)
Post=1 × Treated=1	0.089*** (0.011)	0.084*** (0.011)	0.064*** (0.011)	0.014 (0.012)
Year Trend	Y	Y	Y	
Year FE	Y	Y	Y	
Year of Study	Y	Y	Y	Y
Exams	590039	512494	368806	148987
Students	19573	19426	19070	17725

The dependent variable is standardized course grade. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014 in first column, Fall 2009 through Winter 2013 in second column, and so on. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 4: Effect of OTG on Academic Probation

	(1)	(2)	(3)	(4)	(5)
	Probation	Probation	Probation	Probation	Probation
Post=1	-0.039*** (0.006)	0.026*** (0.006)	0.021*** (0.005)	0.021*** (0.005)	0.023*** (0.006)
Post=1 × Treated=1	-0.045*** (0.006)	-0.029*** (0.006)	-0.029*** (0.006)	-0.029*** (0.006)	-0.030*** (0.006)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Mean of Dep. Var.	.221	.221	.221	.221	.221
Exams	650202	650202	650202	650202	650202
Students	19597	19597	19597	19597	19597

The dependent variable is whether a student is on academic probation (cumulative GPA below C). Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 5: Effect of OTG on Fail Probability

	(1)	(2)	(3)	(4)	(5)
	Fail	Fail	Fail	Fail	Fail
Post=1	-0.004** (0.002)	0.008*** (0.002)	0.008*** (0.002)	0.008*** (0.002)	0.004** (0.002)
Post=1 × Treated=1	-0.010*** (0.002)	-0.007*** (0.002)	-0.007*** (0.002)	-0.007*** (0.002)	-0.007*** (0.002)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Mean of Dep. Var.	.033	.033	.033	.033	.033
Exams	650202	650202	650202	650202	650202
Students	19597	19597	19597	19597	19597

The dependent variable is whether a student fails a course (below a D and no degree progression accumulated). Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 6: Heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)
	Z-Score	Z-Score	Probation	Probation	Fail	Fail
Post=1	-0.094*** (0.018)	-0.076*** (0.011)	0.019* (0.010)	0.021*** (0.006)	0.009*** (0.003)	0.005** (0.002)
Post=1 × Treated=1	0.086*** (0.020)	0.074*** (0.012)	-0.029** (0.011)	-0.020*** (0.007)	-0.010*** (0.003)	-0.005*** (0.002)
Post=1 × Female=1	0.028 (0.022)		0.006 (0.012)		-0.007** (0.004)	
Post=1 × Treated=1 × Female=1	0.007 (0.024)		-0.002 (0.013)		0.004 (0.004)	
Post=1 × STEM=1		-0.000 (0.028)		0.005 (0.013)		-0.001 (0.005)
Post=1 × Treated=1 × STEM=1		0.057* (0.030)		-0.041*** (0.014)		-0.005 (0.005)
Year Trend	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Year of Study	Y	Y	Y	Y	Y	Y
Exams	590039	590039	650202	650202	650202	650202
Students	19573	19573	19597	19597	19597	19597

The dependent variable is standardized course grade, probation or failed course. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses in column 1 and 2. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 7: Effect of OTG on Course Grades by Income

	(1)	(2)	(3)	(4)	(5)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score
Post=1 × Lower (<50k)	0.130*** (0.014)	-0.047*** (0.014)	-0.068*** (0.014)	-0.068*** (0.014)	-0.065*** (0.014)
Post=1 × Upper (50k+)	0.107*** (0.014)	-0.075*** (0.014)	-0.095*** (0.014)	-0.095*** (0.014)	-0.092*** (0.014)
Post=1 × Treated=1 × Lower (<50k)	0.118*** (0.016)	0.070*** (0.015)	0.069*** (0.015)	0.069*** (0.015)	0.073*** (0.016)
Post=1 × Treated=1 × Upper (50k+)	0.153*** (0.016)	0.107*** (0.015)	0.106*** (0.015)	0.106*** (0.015)	0.109*** (0.015)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
p-value of difference	.105	.08	.085	.085	.089
Exams	578424	578424	578424	578424	578424
Students	19229	19229	19229	19229	19229

Sample split by median income, corresponding to 50,000 CAD. The dependent variable is standardized course grade. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses. Only students with valid postal code. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 8: Effect of OTG on Treatment Group Concentration

	(1)	(2)	(3)	(4)	(5)
	Concentration	Concentration	Concentration	Concentration	Concentration
Post=1	-0.010*** (0.002)	-0.008*** (0.002)	-0.008*** (0.002)	-0.008*** (0.002)	-0.012*** (0.002)
Post=1 × Treated=1	0.010*** (0.002)	0.010*** (0.002)	0.010*** (0.002)	0.010*** (0.002)	0.016*** (0.002)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Mean of Dep. Var.	.895	.895	.895	.895	.895
Exams	650202	650202	650202	650202	650202
Students	19597	19597	19597	19597	19597

The dependent variable is percent of treated group students in the classroom (1 if all treated, 0 if all control). Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table 9: Effect of OTG and Concentration on Course Grades

	(1)	(2)	(3)	(4)	(5)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score
Concentration	-0.148*** (0.024)	-0.126*** (0.024)	-0.124*** (0.024)	-0.124*** (0.024)	-0.129*** (0.024)
Treated=1 × Concentration	0.397*** (0.029)	0.394*** (0.029)	0.386*** (0.029)	0.386*** (0.029)	0.375*** (0.029)
Post=1	0.119*** (0.010)	-0.060*** (0.010)	-0.081*** (0.010)	-0.081*** (0.010)	-0.077*** (0.010)
Post=1 × Treated=1	0.136*** (0.011)	0.088*** (0.011)	0.087*** (0.011)	0.087*** (0.011)	0.089*** (0.011)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Exams	590039	590039	590039	590039	590039
Students	19573	19573	19573	19573	19573

The dependent variable is standardized course grade. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 10: Effect of OTG on Part-Time Study

	(1)	(2)	(3)	(4)	(5)
	Part-Time	Part-Time	Part-Time	Part-Time	Part-Time
Post=1	0.101*** (0.006)	0.015*** (0.006)	0.016*** (0.006)	0.016*** (0.006)	0.012* (0.006)
Post=1 × Treated=1	-0.001 (0.006)	-0.023*** (0.006)	-0.023*** (0.006)	-0.023*** (0.006)	-0.014** (0.006)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Mean of Dep. Var.	.081	.081	.081	.081	.081
Semesters	140349	140349	140349	140349	140349
Students	19597	19597	19597	19597	19597

The unit of observation is the semester. The dependent variable is whether a semester is taken as a reduced academic load (two or less courses). Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 11: Effect of OTG on Summer-Study

	(1)	(2)	(3)	(4)	(5)
	Summer	Summer	Summer	Summer	Summer
Post=1	0.062*** (0.003)	0.076*** (0.003)	0.108*** (0.003)	0.108*** (0.003)	0.067*** (0.003)
Post=1 × Treated=1	-0.010*** (0.003)	-0.007** (0.003)	-0.005* (0.003)	-0.005* (0.003)	-0.002 (0.004)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Mean of Dep. Var.	.057	.057	.057	.057	.057
Courses	701260	701260	701260	701260	701260
Students	20104	20104	20104	20104	20104

The dependent variable is whether a course is taken during the summer. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 12: Effect of OTG on Course Grades by Season

	(1)	(2)	(3)	(4)	(5)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score
Post=1	0.131*** (0.013)	-0.085*** (0.013)	-0.108*** (0.013)	-0.108*** (0.013)	-0.116*** (0.013)
Post=1 × Treated=1	0.120*** (0.014)	0.079*** (0.013)	0.076*** (0.013)	0.076*** (0.013)	0.082*** (0.013)
Winter=1	-0.055*** (0.009)	0.001 (0.009)	0.017* (0.009)	0.017* (0.009)	0.033*** (0.009)
Post=1 × Winter=1	0.019 (0.013)	0.035*** (0.013)	0.019 (0.013)	0.019 (0.013)	0.020 (0.013)
Treated=1 × Winter=1	-0.031*** (0.010)	-0.028*** (0.010)	-0.031*** (0.010)	-0.031*** (0.010)	-0.040*** (0.010)
Post=1 × Treated=1 × Winter=1	0.032** (0.014)	0.027* (0.014)	0.031** (0.014)	0.031** (0.014)	0.033** (0.014)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Exams	590039	590039	590039	590039	590039
Students	19573	19573	19573	19573	19573

The dependent variable is standardized course grade. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.

Table 13: Effect of OTG on Graduation Rate

	(1)	(2)	(3)	(4)
	Graduated	Graduated	Graduated	Graduated
Post=1	-0.103*** (0.028)	-0.164** (0.072)	0.105*** (0.030)	-0.164** (0.072)
Post=1 × Treated=1	-0.063** (0.029)	-0.062** (0.028)	-0.055* (0.029)	-0.062** (0.028)
Year Trend			Y	Y
Year FE		Y		Y
Mean of Dep. Var.	.664	.664	.664	.664
Students	25767	25767	25767	25767

The dependent variable is an indicator if a student graduated by the end of the data availability in 2019. Post is an indicator variable for students enrolling after the programs announcement. Treated students from Ontario. Control students from Quebec. Domestic students aged 18 and below. Enrolled as full time only. Enrolled from Fall 2008 through Fall 2014. Heteroskedasticity robust standard errors are in parentheses. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 14: Effect of OTG on Admission Averages

	(1)	(2)	(3)	(4)
	Adm. Avg.	Adm. Avg.	Adm. Avg.	Adm. Avg.
Post=1	0.042** (0.021)	-0.184*** (0.071)	0.112*** (0.023)	-0.184*** (0.071)
Post=1 × Treated=1	-0.039* (0.022)	-0.041* (0.022)	-0.037* (0.022)	-0.041* (0.022)
Year Trend			Y	Y
Year FE		Y		Y
Mean of Dep. Var.	.776	.776	.776	.776
Students	25767	25767	25767	25767

The dependent variable is student admission average. Post is an indicator variable for students enrolling after the programs announcement. Treated students from Ontario. Control students from Quebec. Domestic students aged 18 and below. Enrolled as full time only. Enrolled from Fall 2008 through Fall 2014. Heteroskedasticity robust standard errors are in parentheses. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table 15: Effect of OTG on Persistence

	(1)	(2)	(3)	(4)
	Persisted	Persisted	Persisted	Persisted
Post=1	0.025 (0.020)	-0.077 (0.067)	0.023 (0.021)	-0.077 (0.067)
Post=1 × Treated=1	-0.021 (0.021)	-0.022 (0.021)	-0.021 (0.021)	-0.022 (0.021)
Year Trend			Y	Y
Year FE		Y		Y
Mean of Dep. Var.	.893	.893	.893	.893
Students	25180	25180	25180	25180

The dependent variable is an indicator if a student persisted beyond first year. Post is an indicator variable for students enrolling after the programs announcement. Treated students from Ontario. Control students from Quebec. Domestic students aged 18 and below. Enrolled as full time only. Enrolled from Fall 2008 through Fall 2014. Heteroskedasticity robust standard errors are in parentheses. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

3.8 Appendices

Table A1: Effect of OTG on Course Grades (Local Students)

	(1)	(2)	(3)	(4)	(5)
	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score
Post=1	0.141*** (0.014)	-0.043*** (0.014)	-0.066*** (0.014)	-0.066*** (0.014)	-0.063*** (0.014)
Post=1 × Treated=1	0.113*** (0.015)	0.072*** (0.014)	0.071*** (0.014)	0.071*** (0.014)	0.075*** (0.014)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Exams	336567	336567	336567	336567	336567
Students	11428	11428	11428	11428	11428

Students within 50km only. The dependent variable is standardized course grade. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table A2: Assigning Letter Grades to Percent

Letter Grade	Percentage Interval	Assigned Value
A+	90-100	95
A	85-89	87
A-	80-84	82
B+	75-79	77
B	70-74	72
C+	65-69	67
C	60-64	62
D+	55-59	57
D	50-54	52
E	40-49	44.5
F	0-39	19.5

Student performance at the course level is reported as one of 10 letter grades (first column) which correspond to percentage intervals (second column). The assigned value of a letter grade, which is later standardized is in the third column.

Table A3: Effect of OTG on Student Address Income

	(1)	(2)	(3)	(4)	(5)
	Income	Income	Income	Income	Income
Post=1	-0.037 (0.081)	-0.047 (0.085)	-0.051 (0.083)	-0.051 (0.083)	-0.035 (0.085)
Post=1 × Treated=1	0.073 (0.094)	0.071 (0.092)	0.072 (0.092)	0.072 (0.092)	0.051 (0.094)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Mean of Dep. Var.	52.834	52.834	52.834	52.834	52.834
Courses	637328	637328	637328	637328	637328
Students	19251	19251	19251	19251	19251

The dependent variable is student address average income. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table A4: Effect of OTG on Student Enrollment Address Income

	(1)	(2)	(3)	(4)
	Income	Income	Income	Income
Post=1	-1.855	-5.459**	-1.976	-5.459**
	(1.284)	(2.556)	(1.343)	(2.556)
Post=1 × Treated=1	1.890	1.810	1.886	1.810
	(1.310)	(1.313)	(1.310)	(1.313)
Year Trend			Y	Y
Year FE		Y		Y
Mean of Dep. Var.	53.235	53.235	53.235	53.235
Students	25253	25253	25253	25253

The dependent variable is student enrollment address average income. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table A5: Effect of OTG on Distance

	(1)	(2)	(3)	(4)	(5)
	Distance	Distance	Distance	Distance	Distance
Post=1	-0.029 (0.022)	-0.013 (0.023)	-0.019 (0.024)	-0.019 (0.024)	-0.018 (0.024)
Post=1 × Treated=1	0.014 (0.027)	0.018 (0.027)	0.017 (0.027)	0.017 (0.027)	0.013 (0.027)
Year Trend		Y		Y	Y
Year FE			Y	Y	Y
Year of Study					Y
Mean of Dep. Var.	11.446	11.446	11.446	11.446	11.446
Courses	323105	323105	323105	323105	323105
Students	10128	10128	10128	10128	10128

The dependent variable is distance to student address. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table A6: Effect of OTG on Enrollment Distance

	(1)	(2)	(3)	(4)
	Distance	Distance	Distance	Distance
Post=1	-0.329 (0.289)	-3.279*** (1.046)	-0.256 (0.339)	-3.279*** (1.046)
Post=1 × Treated=1	0.362 (0.310)	0.289 (0.313)	0.363 (0.310)	0.289 (0.313)
Year Trend			Y	Y
Year FE		Y		Y
Mean of Dep. Var.	11.948	11.948	11.948	11.948
Students	12177	12177	12177	12177

The dependent variable is distance to student enrollment address. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table A7: Effect of OTG on Course Grades, Alternative Clustering Strategies

	(1)	(2)	(3)	(4)	(5)
	Unclustered	Preferred - Student	First Semester	Cohort	Coh. \times Prov.
Post=1	-0.077*** (0.008)	-0.077*** (0.010)	-0.077*** (0.014)	-0.077*** (0.013)	-0.077*** (0.020)
Post=1 \times Treated=1	0.089*** (0.008)	0.089*** (0.011)	0.089** (0.040)	0.089*** (0.024)	0.089*** (0.023)
Year Trend	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Year of Study	Y	Y	Y	Y	Y
Exams	590039	590039	590039	590039	590039
Students	19573	19573	19573	19573	19573
Clusters		19573	10	5	10

The dependent variable is standardized course grade. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Homoskedastic errors in the first column. Heteroskedasticity robust standard errors are clustered at the student level in the second column. Clustering at the first semester and cohort levels in the third and fourth column, respectively. In the fifth column, I cluster by province \times cohort. No pass/fail courses. (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.)

Table A8: Effect of OTG on Course Grades, Alternative Standardizations

	(1)	(2)	(3)	(4)
	Z-Score	Z-Score	Z-Score	Score (%)
Post=1	-0.077*** (0.010)	-0.041*** (0.010)	-0.073*** (0.010)	-1.057*** (0.144)
Post=1 × Treated=1	0.089*** (0.011)	0.075*** (0.011)	0.085*** (0.010)	1.230*** (0.151)
Year Trend	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Year of Study	Y	Y	Y	Y
Exams	590039	650202	590039	590039
Students	19573	19597	19573	19573

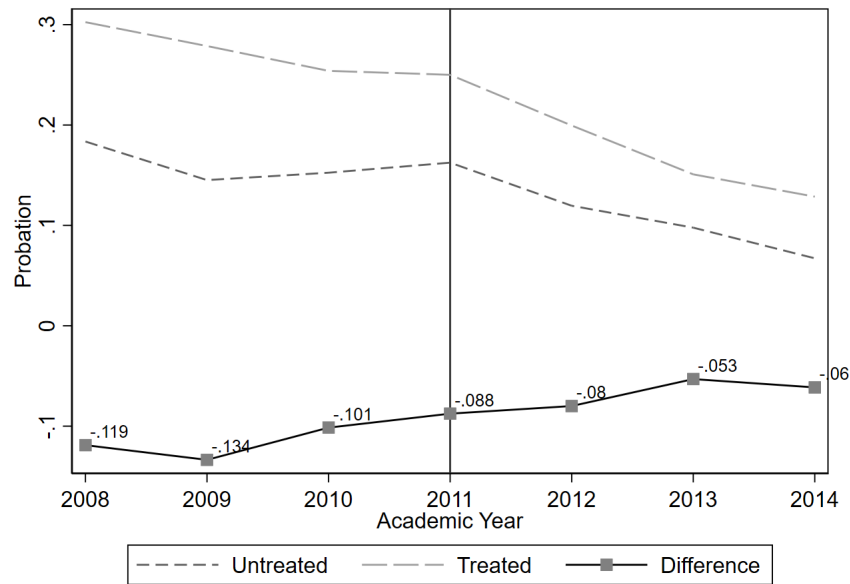
In the first column, the dependent variable is standardized course grade between all treated and control switchers and all years, as used throughout. In the second column, standardization is within treatment group and all years. In the third column, standardization is between all students (including, for example, foreign students). In the fourth column, no standardization is applied. Post is an indicator variable for after and including Winter 2012. Treated students from Ontario. Control students from Quebec. Within-student fixed effects model. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Exams from Fall 2008 through Winter 2014. Heteroskedasticity robust standard errors are in parentheses, clustered at the student level. No pass/fail courses. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Table A9: Effect of OTG on Dropout by Season

	(1)	(2)	(3)	(4)
	Winter Dropout	Winter Dropout	Winter Dropout	Winter Dropout
Post=1	-0.069 (0.044)	0.105 (0.080)	0.006 (0.047)	0.105 (0.080)
Post=1 × Treated=1	-0.010 (0.045)	-0.002 (0.045)	-0.004 (0.045)	-0.002 (0.045)
Year Trend			Y	Y
Year FE		Y		Y
Mean of Dep. Var.	.751	.751	.751	.751
Students	7773	7773	7773	7773

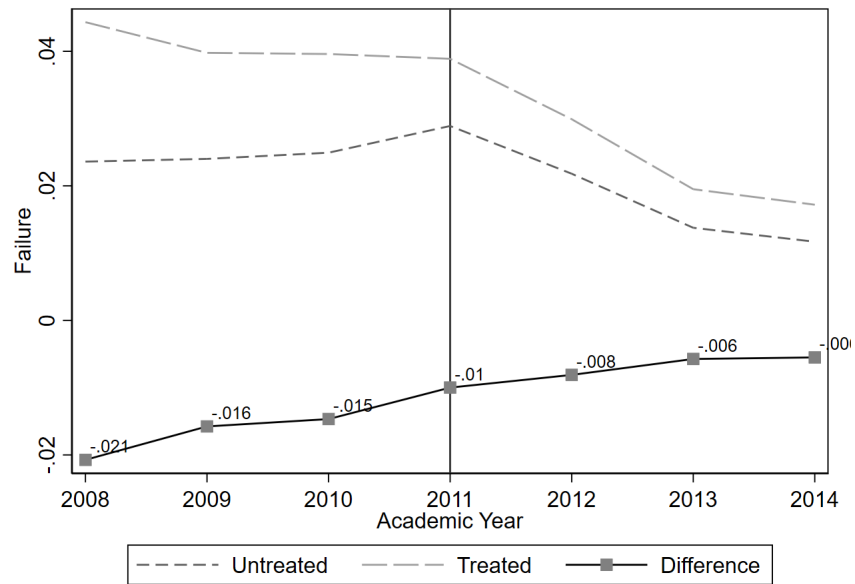
The dependent variable is an indicator that takes a value one if a student dropped out in the winter term. It takes the value of zero if the student dropped out in the fall term. Post is an indicator variable for students enrolling after the programs announcement. Treated students from Ontario. Control students from Quebec. Domestic students aged 18 and below. Enrolled as full time only. Enrolled from Fall 2008 through Fall 2014. Heteroskedasticity robust standard errors are in parentheses. (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.)

Figure A1: Trends by Treatment Group - Probation



Percent of course grades taken under academic probation by academic year. Treated students from Ontario. Control students from Quebec. Difference plotted as a solid line with values attached. Vertical line in 2011 provided when the Grant was announced and introduced. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Fall 2008 through Winter 2014.

Figure A2: Trends by Treatment Group - Failure



Percent of courses failed by academic year. Treated students from Ontario. Control students from Quebec. Difference plotted as a solid line with values attached. Vertical line in 2011 provided when the Grant was announced and introduced. Students who move from post=0 to post=1. Domestic students aged 22 and below. Enrolled as full time only. Fall 2008 through Winter 2014.

Bibliography

- Joseph Altonji and Rebecca Blank. Race and gender in the labor market. In O. Ashenfelter and D. Card, editors, *Handbook of Labor Economics*, volume 3, Part C, chapter 48, pages 3143–3259. Elsevier, 1 edition, 1999.
- Ofra Amir, David G Rand, et al. Economic games on the internet: The effect of \$1 stakes. *PloS one*, 7(2):e31461, 2012.
- Joshua Angrist, Daniel Lang, and Philip Oreopoulos. Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–63, 2009.
- Kenneth J Arrow. What has economics to say about racial discrimination? *Journal of economic perspectives*, 12(2):91–100, 1998.
- LE Banderet, DM MacDougall, DE Roberts, D Tappan, and M Jacey. Effects of hypohydration or cold exposure and restricted fluid intake upon cognitive performance. Technical report, United States Army Research Institute of Environmental Medicine, 1986.
- Nick Bansback, John Brazier, Aki Tsuchiya, and Aslam Anis. Using a discrete choice experiment to estimate health state utility values. *Journal of health economics*, 31(1):306–318, 2012.
- Andrew Barr. Fighting for education: Financial aid and degree attainment. *Journal of Labor Economics*, 37(2):509–544, 2019.
- Alan Barreca, Karen Clay, Olivier Deschenes, Michael Greenstone, and Joseph S Shapiro. Adapting to climate change: The remarkable decline in the US temperature-mortality relationship over the twentieth century. *Journal of Political Economy*, 124(1):105–159, 2016.
- Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800, 2011.
- DOUGLAS G Bell, PETER Tikuisis, and I Jacobs. Relative intensity of muscular contraction during shivering. *Journal of Applied Physiology*, 72(6):2336–2342, 1992.
- Marianne Bertrand and Esther Duflo. Field experiments on discrimination. In *Handbook of Economic Field Experiments*, volume 1, pages 309–393. Elsevier, 2017.

- Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The American Economic Review*, 94(4):991–1013, 2004.
- Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.
- Marianne Bertrand, Dolly Chugh, and Sendhil Mullainathan. Implicit discrimination. *American Economic Review*, 95(2):94–98, 2005.
- Timothy Besley and Robin Burgess. Can labor regulation hinder economic performance? evidence from india. *The Quarterly journal of economics*, 119(1):91–134, 2004.
- Eric Bettinger, Oded Gurantz, Laura Kawano, Bruce Sacerdote, and Michael Stevens. The long-run impacts of financial aid: Evidence from california’s cal grant. *American Economic Journal: Economic Policy*, 11(1):64–94, 2019.
- Hoyt Bleakley. Malaria eradication in the Americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics*, 2(2):1–45, 2010.
- Zachary Bleemer and Basit Zafar. Intended college attendance: Evidence from an experiment on college returns and costs. *Journal of Public Economics*, 157:184–211, 2018.
- Denis P Blondin, Sébastien M Labbé, Hans C Tingelstad, Christophe Noll, Margaret Kunach, Serge Phoenix, Brigitte Guérin, Éric E Turcotte, André C Carpentier, Denis Richard, et al. Increased brown adipose tissue oxidative capacity in cold-acclimated humans. *The Journal of Clinical Endocrinology & Metabolism*, 99(3):E438–E446, 2014.
- Maarten AS Boksem, Theo F Meijman, and Monique M Lorist. Effects of mental fatigue on attention: an erp study. *Cognitive brain research*, 25(1):107–116, 2005.
- Marius Brazaitis, Nerijus Eimantas, Laura Daniuseviciute, Dalia Mickeviciene, Rasa Steponaviciute, and Albertas Skurvydas. Two strategies for response to 14 c cold-water immersion: is there a difference in the response of motor, cognitive, immune and stress markers? *PLoS One*, 9(10):e109020, 2014.
- Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.
- Robin Burgess, Olivier Deschenes, Dave Donaldson, and Michael Greenstone. Weather, climate change and death in india. 2017.
- A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.
- David Card, Martin D Dooley, and A Abigail Payne. School competition and efficiency with publicly funded catholic schools. *American Economic Journal: Applied Economics*, 2(4):150–76, 2010.

- Deven E Carlson, Felix Elwert, Nicholas Hillman, Alex Schmidt, and Barbara L Wolfe. The effects of financial aid grant offers on postsecondary educational outcomes: New experimental evidence from the fund for wisconsin scholars. Technical report, National Bureau of Economic Research, 2019.
- Benjamin L Castleman and Bridget Terry Long. Looking beyond enrollment: The causal effect of need-based grants on college access, persistence, and graduation. *Journal of Labor Economics*, 34(4):1023–1073, 2016.
- Gary Charness and Uri Gneezy. What’s in a name? anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, 68(1):29–35, 2008.
- Gary Charness, Luca Rigotti, and Aldo Rustichini. Individual behavior and group membership. *American Economic Review*, 97(4):1340–1352, 2007.
- Daniel Chen, Yosh Halberstam, and CL Alan. Perceived masculinity predicts us supreme court outcomes. *PLoS ONE*, 11(10):e0164324, 2016.
- Jiu-Chiuan Chen and Joel Schwartz. Neurobehavioral effects of ambient air pollution on cognitive performance in US adults. *Neurotoxicology*, 30(2):231–239, 2009.
- Roy Chen and Yan Chen. The potential of social identity for equilibrium selection. *American Economic Review*, 101(6):2562–89, 2011.
- Stephen S Cheung, Jason KW Lee, and Juha Oksa. Thermal stress, human performance, and physical employment standards. *Applied Physiology, Nutrition, and Metabolism*, 41(6):S148–S164, 2016.
- Scott Clifford, Ryan M Jewell, and Philip D Waggoner. Are samples drawn from mechanical turk valid for research on political ideology? *Research & Politics*, 2(4):2053168015622072, 2015.
- Judah Cohen, Karl Pfeiffer, and Jennifer A Francis. Warm arctic episodes linked with increased frequency of extreme winter weather in the united states. *Nature communications*, 9(1):869, 2018.
- Michael R Cunningham. Weather, mood, and helping behavior: Quasi experiments with the sunshine samaritan. *Journal of Personality and Social Psychology*, 37(11):1947, 1979.
- Hein AM Daanen and Wouter D Van Marken Lichtenbelt. Human whole body cold adaptation. *Temperature*, 3(1):104–118, 2016.
- Hein AM Daanen, Evert Van De Vliert, and Xu Huang. Driving performance in cold, warm, and thermonneutral environments. *Applied Ergonomics*, 34(6):597–602, 2003.
- Esther W de Bekker-Grob, Bas Donkers, Marcel F Jonker, and Elly A Stolk. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. *The Patient-Patient-Centered Outcomes Research*, 8(5):373–384, 2015. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4575371/#CR37>.

- Melissa Dell, Benjamin F Jones, and Benjamin A Olken. Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*, 4(3):66–95, 2012.
- Stefano DellaVigna and Devin Pope. What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 2016.
- Jeffrey T Denning. College on the cheap: Consequences of community college tuition reductions. *American Economic Journal: Economic Policy*, 9(2):155–88, 2017.
- Jeffrey T Denning. Born under a lucky star financial aid, college completion, labor supply, and credit constraints. *Journal of Human Resources*, 54(3):760–784, 2019.
- Tatyana Deryugina and Solomon M Hsiang. Does the environment still matter? daily temperature and income in the United States. Technical Report 20750, National Bureau of Economic Research, 2014.
- GC Donaldson, H Rintamäki, and S Näyhä. Outdoor clothing: its relationship to geography, climate, behaviour and cold-related mortality in europe. *International Journal of Biometeorology*, 45(1):45–51, 2001.
- Susan Dynarski and Judith Scott-Clayton. Financial aid policy: Lessons from research. Technical report, National Bureau of Economic Research, 2013.
- Susan Dynarski and Mark Wiederspan. Student aid simplification: Looking back and looking ahead. Technical report, National Bureau of Economic Research, 2012.
- Susan M Dynarski. Does aid matter? measuring the effect of student aid on college attendance and completion. *American Economic Review*, 93(1):279–288, 2003.
- Avraham Ebenstein, Victor Lavy, and Sefi Roth. The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution. *American Economic Journal: Applied Economics*, 8(4):36–65, 2016.
- Catherine C Eckel and Ragan Petrie. Face value. *American Economic Review*, 101(4):1497–1513, 2011.
- Ann Enander. Performance and sensory aspects of work in cold environments: A review. *Ergonomics*, 27(4):365–378, 1984.
- Ann Enander, Björn Sköldström, and Ingvar Holmér. Reactions to hand cooling in workers occupationally exposed to cold. *Scandinavian Journal of Work, Environment & Health*, pages 58–65, 1980.
- Niklas Engbom and Christian Moser. Returns to education through access to higher-paying firms: Evidence from us matched employer-employee data. *American Economic Review*, 107(5):374–78, 2017.
- Tor Eriksson and Nicolai Kristensen. Wages or fringes? some evidence on trade-offs and sorting. *Journal of Labor Economics*, 32(4):899–928, 2014.

- Léon G Faber, Natasha M Maurits, and Monique M Lorst. Mental fatigue affects visual selective attention. *PloS one*, 7(10):e48073, 2012.
- Jan Feld, Nicolás Salamanca, and Daniel Hamermesh. Endophilia or exophobia: Beyond discrimination. *The Economic Journal*, 126(August):1503–1527, 2015.
- David Figlio, Krzysztof Karbownik, and Kjell G Salvanes. Education research and administrative data. In *Handbook of the Economics of Education*, volume 5, pages 75–138. Elsevier, 2016.
- Bernard J Fine. The effect of exposure to an extreme stimulus on judgments of some stimulus-related words. *Journal of Applied Psychology*, 45(1):41, 1961.
- Adam Finn and Jordan J Louviere. Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing*, pages 12–25, 1992.
- Ross Finnie and Richard Mueller. Access to post-secondary education: How does québec compare to the rest of canada? *L'Actualité économique*, 93(3):441–474, 2017.
- Robert J Fisher. Social desirability bias and the validity of indirect questioning. *Journal of consumer research*, 20(2):303–315, 1993.
- Jeffrey A Flory, Andreas Leibbrandt, and John A List. Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82(1):122–155, 2014.
- Reuben Ford, Taylor Shek-wai Hui, and Cam Nguyen. *Postsecondary Participation and Household Income*. Higher Education Quality Council of Ontario, 2019.
- Lorenz Goette, David Huffman, and Stephan Meier. The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics*, 4(1):101–15, 2012.
- Joshua Goodman, Michael Hurwitz, Jisung Park, and Jonathan Smith. Heat and learning. *American Economic Journal: Economic Policy*, Forthcoming.
- Joshua Graff Zivin and Matthew Neidell. Temperature and the allocation of time: Implications for climate change. *Journal of Labor Economics*, 32(1):1–26, 2014.
- Joshua Graff Zivin, Solomon M Hsiang, and Matthew Neidell. Temperature and human capital in the short and long run. *Journal of the Association of Environmental and Resource Economists*, 5(1):77–105, 2018.
- Daniel S Hamermesh, Jeff E Biddle, et al. Beauty and the labor market. *American Economic Review*, 84(5):1174–1194, 1994.
- Nick Hanley, Robert E Wright, and Vic Adamowicz. Using choice experiments to value the environment. *Environmental and resource economics*, 11(3-4):413–428, 1998.

- James Hansen, Makiko Sato, and Reto Ruedy. Perception of climate change. *Proceedings of the National Academy of Sciences*, 109(37):E2415–E2423, 2012.
- GH Hartung, LG Myhre, and SA Nunneley. Physiological effects of cold air inhalation during exercise. *Aviation, Space, and Environmental Medicine*, 51(6):591–594, 1980.
- Jerry Hausman and Daniel McFadden. Specification tests for the multinomial logit model. *Econometrica: Journal of the Econometric Society*, pages 1219–1240, 1984.
- Morten S. Hedegaard and Jean-Robert Tyran. The price of prejudice. *American Economic Journal: Applied Economics*, 10(1):40–63, 2018.
- Anthony Heyes and John A List. Supply and demand for discrimination: strategic revelation of own characteristics in a trust game. *American Economic Review*, 106(5):319–23, 2016.
- Anthony Heyes and Soodeh Saberian. Temperature and decisions: Evidence from 207,000 court cases. *American Economic Journal: Applied Economics*, 11(2):238–65, 2019.
- John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425, 2011.
- Connor Huff and Dustin Tingley. “who are these people?” evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):2053168015604648, 2015.
- Richard M Johnson and Bryan K Orme. How many questions should you ask in choice-based conjoint studies. In *Sawtooth Software Research Paper Series*, 1996.
- Robert J Johnston, Kevin J Boyle, Wiktor Adamowicz, Jeff Bennett, Roy Brouwer, Trudy Ann Cameron, W Michael Hanemann, Nick Hanley, Mandy Ryan, Riccardo Scarpa, et al. Contemporary guidance for stated preference studies. *Journal of the Association of Environmental and Resource Economists*, 4(2):319–405, 2017.
- Ethel B Jones and John D Jackson. College grades and labor market rewards. *The Journal of Human Resources*, 25(2):253–266, 1990.
- Melissa S Kearney and Phillip B Levine. Media influences on social outcomes: The impact of mtv’s 16 and pregnant on teen childbearing. *American Economic Review*, 105(12):3597–3632, 2015.
- Jeremy Kees, Christopher Berry, Scot Burton, and Kim Sheehan. An analysis of data quality: professional panels, student subject pools, and amazon’s mechanical turk. *Journal of Advertising*, 46(1):141–155, 2017.
- Melissa G Keith, Louis Tay, and Peter D Harms. Systems perspective of amazon mechanical turk for organizational research: Review and recommendations. *Frontiers in psychology*, 8:1359, 2017.

- Baek-Min Kim, Seok-Woo Son, Seung-Ki Min, Jee-Hoon Jeong, Seong-Joong Kim, Xiangdong Zhang, Taehyoun Shim, and Jin-Ho Yoon. Weakening of the stratospheric polar vortex by arctic sea-ice loss. *Nature Communications*, 5:4646, 2014.
- Hyoun S Kim and David C Hodgins. Reliability and validity of data obtained from alcohol, cannabis, and gambling populations on amazon's mechanical turk. *Psychology of addictive behaviors*, 31(1):85, 2017.
- Brian Knight and Nathan Schiff. The out-of-state tuition distortion. *American Economic Journal: Economic Policy*, 11(1):317–50, 2019.
- Erik W Kolstad, Tarjei Breiteig, and Adam A Scaife. The association between stratospheric weak polar vortex events and cold air outbreaks in the Northern Hemisphere. *Quarterly Journal of the Royal Meteorological Society*, 136(649):886–893, 2010.
- Marlene Kretschmer, Dim Coumou, Laurie Agel, Mathew Barlow, Eli Tziperman, and Judah Cohen. More-persistent weak stratospheric polar vortex states linked to cold extremes. *Bulletin of the American Meteorological Society*, 99(1):49–60, 2018.
- Jonathan Kropko. *Choosing between multinomial logit and multinomial probit models for analysis of unordered choice data*. PhD thesis, The University of North Carolina at Chapel Hill, 2007.
- Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236, 1991.
- Peter Kuhn and Marie Claire Villeval. Are women more attracted to co-operation than men? *The Economic Journal*, 125(582):115–140, 2015.
- Ilyana Kuziemko, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva. How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review*, 105(4):1478–1508, 2015.
- Tom Lane. Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review*, 90:375–402, 2016.
- Jean-Claude Launay and Gustave Savourey. Cold adaptations. *Industrial Health*, 47(3):221–227, 2009.
- Emily CP LaVoy, Brian K McFarlin, and Richard J Simpson. Immune responses to exercising in a cold environment. *Wilderness & Environmental Medicine*, 22(4):343–351, 2011.
- John Leach, A Abigail Payne, and Steve Chan. The effects of school board consolidation and financing on student performance. *Economics of Education Review*, 29(6):1034–1046, 2010.
- Joa Julia Lee, Francesca Gino, and Bradley R Staats. Rainmakers: Why bad weather means good productivity. *Journal of Applied Psychology*, 99(3):504, 2014.

- S.D. Levitt and S.J. Dubner. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. HarperCollins, 2011. ISBN 9780062132345. URL <https://books.google.ca/books?id=wNPn15zYA-cC>.
- Jason M Lindo, Nicholas J Sanders, and Philip Oreopoulos. Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2(2):95–117, 2010.
- John A List and Fatemeh Momeni. When corporate social responsibility backfires: Theory and evidence from a natural field experiment. Working Paper 24169, National Bureau of Economic Research, December 2017.
- Jordan J Louviere, Terry N Flynn, and Richard T Carson. Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling*, 3(3):57–72, 2010.
- Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavioural Research*, 47:1122–1135, 2015.
- Tiina M Mäkinen. Human cold exposure, adaptation, and performance in high latitude environments. *American Journal of Human Biology*, 19(2):155–164, 2007.
- Day Manoli and Nicholas Turner. Cash-on-hand and college enrollment: Evidence from population tax data and the earned income tax credit. *American Economic Journal: Economic Policy*, 10(2):242–71, 2018.
- Alexandre Mas and Amanda Pallais. Valuing alternative work arrangements. *American Economic Review*, 107(12):3722–3759, 2017.
- Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1974. URL <https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf>.
- Daniel McFadden. The choice theory approach to market research. *Marketing science*, 5(4):275–297, 1986.
- Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, pages 447–470, 2000.
- Markus M Mobius and Tanya S Rosenblat. Why beauty matters. *American Economic Review*, 96(1):222–235, 2006.
- Matthew D Muller, John Gunstad, Michael L Alosco, Lindsay A Miller, John Updegraff, Mary Beth Spitznagel, and Ellen L. Glickman. Acute cold exposure and cognitive function: Evidence for sustained impairment. *Ergonomics*, 55(7):792–798, 2012.
- David Neumark. Experimental research on labor market discrimination. Working Paper 22022, National Bureau of Economic Research, February 2016. URL <http://www.nber.org/papers/w22022>.
- Jennifer L Nguyen, Joel Schwartz, and Douglas W Dockery. The relationship between indoor and outdoor temperature, apparent temperature, relative humidity, and absolute humidity. *Indoor Air*, 24(1):103–112, 2014.

- Clemens Noelke, Mark McGovern, Daniel J Corsi, Marcia P Jimenez, Ari Stern, Ian Sue Wing, and Lisa Berkman. Increasing ambient temperature reduces emotional well-being. *Environmental Research*, 151:124–129, 2016.
- Philip Oreopoulos. Why do skilled immigrants struggle in the labor market? a field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3(4):148–71, 2011.
- Bryan Orme. Sample size issues for conjoint analysis studies. *Sawtooth Software Research paper Series. Squim, WA, USA: Sawtooth Software Inc*, 1998.
- Lindsay C Page and Judith Scott-Clayton. Improving college access in the united states: Barriers and policy responses. *Economics of Education Review*, 51:4–22, 2016.
- Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5), 2010.
- Jisung Park. Hot temperature and high stakes exams: Evidence from new york city public schools. *Unpublished Manuscript, Harvard University Economics Department*, 2016.
- Priya G Patil, Jeffrey L Apfelbaum, and James P Zacny. Effects of a cold-water stressor on psychomotor and cognitive functioning in humans. *Physiology & Behavior*, 58(6):1281–1286, 1995.
- June J Pilcher, Eric Nadler, and Caroline Busch. Effects of hot and cold temperature exposure on performance: A meta-analytic review. *Ergonomics*, 45(10):682–698, 2002.
- Devin G Pope and Justin R Sydnor. What’s in a picture? evidence of discrimination from prosper. com. *Journal of Human Resources*, 46(1):53–92, 2011.
- Ashlinn Quinn, James D Tamerius, Matthew Perzanowski, Judith S Jacobson, Inge Goldstein, Luis Acosta, and Jeffrey Shaman. Predicting indoor heat exposure risk during extreme heat events. *Science of the Total Environment*, 490: 686–693, 2014.
- Dipankar Rai, Chien-Wei Lin, and Chun-Ming Yang. The effects of temperature cues on charitable donation. *Journal of Consumer Marketing*, 34(1):20–28, 2017.
- Sherwin Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974.
- Sherwin Rosen. The theory of equalizing differences. *Handbook of Labor Economics*, 1:641–692, 1986.
- Bradley J Ruffle and Ze’ev Shtudiner. Are good-looking people more employable? *Management Science*, 61(8):1760–1776, 2014.

- Nicholas O Rule and Nalini Ambady. The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological science*, 19(2):109–111, 2008.
- John Karl Scholz and Kamil Sicinski. Facial attractiveness and lifetime earnings: Evidence from a cohort study. *Review of Economics and Statistics*, 97(1):14–28, 2015.
- Joel Schwartz, Jonathan M Samet, and Jonathan A Patz. Hospital admissions for heart disease: The effects of temperature and humidity. *Epidemiology*, 15(6):755–761, 2004.
- Daniel J Simons and Christopher F Chabris. Common (mis) beliefs about memory: A replication and comparison of telephone and mechanical turk survey methods. *PloS one*, 7(12):e51876, 2012.
- E Somanathan, Rohini Somanathan, Anant Sudarshan, Meenu Tewari, et al. The impact of temperature on productivity and labor supply: Evidence from Indian manufacturing. Technical Report 14-10, Indian Statistical Institute, New Delhi, India, 2015.
- JD Tamerius, MS Perzanowski, LM Acosta, JS Jacobson, IF Goldstein, JW Quinn, AG Rundle, and J Shaman. Socioeconomic and outdoor meteorological determinants of indoor temperature and humidity in new york city dwellings. *Weather, Climate, and Society*, 5(2):168–179, 2013.
- Lee Taylor, Samuel L Watkins, Hannah Marshall, Ben J Dascombe, and Josh Foster. The impact of different environmental conditions on cognitive function: A focused review. *Frontiers in Physiology*, 6:372, 2016.
- Warren H Teichner. Reaction time in the cold. *Journal of Applied Psychology*, 42(1):54, 1958.
- John R Thomas, Stephen T Ahlers, John F House, and John Schrot. Repeated exposure to moderate cold impairs matching-to-sample performance. *Aviation, Space, and Environmental Medicine*, 1989.
- Yutaka Tochihara. Work in artificial cold environments. *Journal of Physiological Anthropology and Applied Human Science*, 24(1):73–76, 2005.
- Alexander Todorov, Anesu N Mandisodza, Amir Goren, and Crystal C Hall. Inferences of competence from faces predict election outcomes. *Science*, 308(5728):1623–1626, 2005.
- Samuel L Watkins, Paul Castle, Alexis R Mauger, Nicholas Sculthorpe, Natalie Fitch, Jeffrey Aldous, John Brewer, Adrian W Midgley, and Lee Taylor. The effect of different environmental conditions on the decision-making performance of soccer goal line officials. *Research in Sports Medicine*, 22(4):425–437, 2014.
- Matthew Wiswall and Basit Zafar. Preference for the workplace, investment in human capital and gender. *The Quarterly Journal of Economics*, 133(1):457–507, 2016.
- Michael Yeomans, Anuj K Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Management Science*, 2017.

Peng Zhang, Olivier Deschenes, Kyle Meng, and Junjie Zhang. Temperature effects on productivity and factor reallocation: Evidence from a half million chinese manufacturing plants. *Journal of Environmental Economics and Management*, 88: 1–17, 2018.

Joshua S Graff Zivin, Yingquan Song, Qu Tang, and Peng Zhang. Temperature and high-stakes cognitive performance: evidence from the national college entrance examination in china. Technical report, National Bureau of Economic Research, 2018.