UNIVERSITY OF CALIFORNIA

Santa Barbara

Essays on Labor Economics and Econometrics

A Dissertation submitted in partial satisfaction of the

requirements for the Degree Doctor of Philosophy

in Economics

by

Benjamin Hansen

Committee in Charge:

Professor Peter J. Kuhn, Co-Chair

Professor Douglas G. Steigerwald, Co-Chair

Professor Olivier Deschênes

June 2009

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

The Dissertation of Benjamin Hansen is approved.

Olivier Deschênes

Douglas G. Steigerwald, Committee Co-Chair

Peter J. Kuhn, Committee Co-Chair

June 2009

Essays on Labor Economics and Econometrics

by

Benjamin Hansen

To Rileigh, Jocelyn, and Brooklyn

I love you girls, you make every day in my life a joy.

Acknowledgments

I want to thank my committee, Peter, Doug, and Olivier for their invaluable comments, advice, and encouragement. I am also grateful to Kelly Bedard, Javier Birchenall, Philip Babcock, Jon Sonstelie, Cathy Weinberger, Heather Royer, Peter Rupert, Patrik Guggenberger, Dave Marcotte and Steven Hemelt for advice which helped improve the quality and direction of my research. Participants at the UCSB Labor Lunch, Libratory Aggregate Economics and Finance, and Econometrics Lab also gave valuable feedback, as did those who attended the AEFA, WEAI, APPAM and SOLE annual meetings. Special thanks to all of those who helped provide me with data.

VITA OF BENJAMIN HANSEN
June 2009

EDUCATION

B.A. in Economics, Brigham Young University April 2004
M.A. in Economics, University of California Santa Barbara, July 2005
Ph.D. in Economics, University of California Santa Barbara, June 2009 (expected)

PROFESSIONAL EMPLOYMENT/EXPERIENCE

2003-2004, Teaching and Research Assistant, Brigham Young University
2005-2009, Teaching Assistant, University of California Santa Barbara
2006-2009, Research Assistant, University of California Santa Barbara
2009- , Research Associate, Impaq International, LLC

CONFERENCE PRESENTATIONS

2009: AEA and SOLE Annual Meetings
2008: AEFA, SOLE, WEAI, and APPAM Annual Meetings, UC-Wide Labor and UCTC Conferences
2007: Causes and Consequences of Earnings Inequality, UC Merced Inaugural Conference

AWARDS

Graduate Opportunity Fellowship, 2008-2009
Distinguished Research Fellowship, 2007
Roe L. Johns Travel Fellowship, 2007
Distinction, Preliminary Examination in Econometrics, 2005
Raymond K. Myerson Academic Excellence Fellowship, 2004

FIELDS OF STUDY

Labor Economics (Under Peter J. Kuhn and Olivier Deschênes)
Economics of Education (Under Peter J. Kuhn and Olivier Deschênes)
Econometrics (Under Douglas G. Steigerwald)

ABSTRACT

Essays on Labor Economics and Econometrics

by

Benjamin Hansen

This dissertation investigates three separate questions relevant to labor economics and econometrics. Chapter 1 focuses on workers' compensation and whether there is evidence supportive of employees' use of workers' compensation to cover off the job injuries. Using administrative data from a large nation-wide temporary staffing agency, I find that difficult-to-diagnose injuries are disproportionately likely to occur on Monday, possibly due to workers being injured off of the job during the weekend and claiming the injury occurred at work on Monday. I also find this phenomena is particularly strong for claims for compensation of lost-wages, rather than medical-only injuries. Focusing on recent reforms in California which increased the level of scrutiny placed on workers' compensation claims, I find a relative decrease in Monday injuries in California only for the difficult-to-diagnose injuries. Taken as a whole, the evidence is consistent with off-the-job injuries composing a fraction of the claims reported on Mondays. Chapter 2 studies the relationship between instructional days and student performance on standardized tests. To identify quasi-random variation in instructional days which

is independent of school inputs, I examine snow-days in Colorado and Maryland and test-date shifts in Minnesota. Both sources of instructional day variation suggest more instructional time improves student performance. Chapter 3 focuses on testing for latent regime-switching, an ongoing challenge in econometrics. We take advantage of recent advances by Cho and White (2007) which proved the existence of a limiting distribution for the likelihood ratio statistic. However, we find the approximation method of Cho and White (2007) requires the specification of the alternative parameter space. If the true alternative – something which is ex-ante unknown to the researcher – lies outside of the pre-specified space, the power of their test can fall to zero for distant alternatives. We find subsampling provides critical values with reasonable size which also negate the drawback of pre-specified coefficient intervals. The power gains are large relative to other established tests for unobserved heterogeneity.

Contents

List of Figures

List of Tables

Part I

Labor Economics

Chapter 1

The Monday Effect in Workers' Compensation: Evidence from the California Reforms

1.1 Introduction

Historically, Monday has been a disproportionately dangerous day to work. Smith (1989) found that the excess fraction of injuries on Monday is largest for injuries that are difficult to diagnose, suggesting that workers might be filing easily-concealed weekend accidents as workers' compensation injuries. If employees are indeed using workers' compensation for weekend injuries, this adverse selection would only add additional externalities to an already costly social insurance program (currently with over $53 billion dollars in annual benefits[1]). While the excess fraction of injuries on Monday is well-documented, if weekend injuries are driving the excess fraction of Monday claims then Monday claiming activity should be sensitive to both the relative benefits of filing false claims. However, previous studies have disagreed on whether benefits affect the likelihood of Monday injuries (see Card and McCall 1996, Ruser 1998 and Campoliete and Hyatt 2006). With that in mind, there remains uncertainty regarding the role weekend injuries play in explaining the prevalence of Monday injuries.

This paper offers the first quasi-experimental evidence to test whether the excess fraction of injuries arising on Monday can be attributed to weekend injuries, examining whether recent reforms in California affected the high incidence of Monday claims. The passage of reforms in California in 2004 allowed employers to choose the doctors rather than the employees, required workers to show what

[1]Source: National Academy of Social Insurance.

fraction of the injury occurred on the job, and limited the duration under which employees could receive temporary total benefits – among many other changes. These recent changes provide a test for the weekend injury hypothesis as the reforms both potentially decrease the expectation of successfully filing false claims and also reduce benefits even if false claims are filed successfully.

Using administrative records and detailed micro data from a large national staffing firm in the United States (employing over 70,000 temporaries on a yearly basis), I investigate injuries and claims in over 35 states (although a plurality are employed in California).[2] Because temporary employment is an industry in which higher levels of asymmetric information can increase moral hazard, the claims data in this analysis provide an ideal setting to link off-the-job injuries to the relative prevalence of Monday injuries.[3] By the nature of the employment situation, it is difficult for the temporary firm to monitor the safety of its workers and also the temporary workers bear little attachment to the firm. In short, if there is no evidence that the excess number of Monday injuries is due to off-the-job injuries in industries like temporary employment, it would be unlikely to be uncovered

[2]There are employees in 41 states, although the hours worked in 6 states is small enough that no injuries are recorded during the 2002-2006 period.

[3]While the growth of temporary employment (see Figure 1 in the Appendix) has attracted the focus of much recent research by economists concerning its effects on earnings and employment, there are many reasons why temporary employment making it a fruitful setting to test for evidence of moral hazard in claiming behavior. Because temporary firms are not present at the job site the asymmetric information between employers and employees can be amplified, which has been evidenced by higher claim rates in temporary employment (Park and Butler 2001). In addition, contingent workers have less job security, an element Fortin and Lanoie (1992) documents can increase claims. Outside of economics, a growing literature addresses these and other concerns regarding the safety of contingent employees (see Virtanen et al. 2005 for an overview).

elsewhere.

Before the reforms, Monday injuries composed 24 % of difficult-to-diagnose injuries and only 19 % of easier-to-diagnose causes. After the reforms, only 17 % of difficult-to-diagnose injuries occurred on Monday in California, with essentially no change for easy-to-diagnose causes. The findings are consistent with the hypothesis that fraudulent claims compose a fraction of the Monday effect, at least in employment settings with substantial asymmetric information between employers and employees – such as temporary employment.

The remainder of the paper proceeds as follows. Section 1.2 provides a background of the literature on workers' compensation, the Monday effect and also explanations for why the reforms in California would impact claiming behavior. Section 1.3 provides initial analysis by demonstrating that prior to the reforms the excess fraction of injuries on Monday was as high as 8.6 percentage points and also showing that even in simple summary statistics the laws appear to have reduced the relative frequency of Monday injuries. Section 1.4 contains a more detailed examination of the effects of the reforms mandated by SB 899 on Monday claims. Section 1.5 summarizes the effect of the reforms on overall claim rates, claim costs (compensation, medical, and legal), thus providing a scale of how important the reduction in Monday claims may be relative to other worker behaviors affected by the reforms. Section 1.6 concludes discussing implications of the results.

1.2 Background

1.2.1 Moral Hazard in Workers' Compensation

As with other forms of social insurance, moral hazard plays a large role in the interaction of potential beneficiaries (employees and employers alike). Regarding workers' compensation, the asymmetric information can largely be decomposed into two forms. First, employers are not able to fully observe the effort that workers exert to avoid injuries – referred to as ex-ante moral hazard. The second, called ex-post moral hazard, concerns the nature and extent of injury which is known by the employee but not to the employer. These two types of informational asymmetries thereby allow workers put forth less safety effort than the firm would desire, exaggerate the injury's severity or misrepresent the cause – conceivably to increase their consumption of leisure.

Along this line, a large literature has found that increased workers' compensation benefits – in the form of either greater replacement rates (the fraction of wages that is replaced with workers' compensation benefits) [4] or shorting waiting periods (the time which must elapse before workers begin to receive compensation benefits) – are associated with increased claims. For instance, Butler and Worrall (1983) and Krueger (1990) find higher benefits increase the number of claims filed, while Meyer et.al. (1995) and Butler and Worrall (1985) associate higher

[4] It is typically 2/3 of gross earnings subject to minimum and maximum thresholds, with benefits received free of taxes. As such both the tax liability of workers and the thresholds create replacement rates often close 80 or 90 percent of after tax earnings.

benefits with longer claim duration. More recently, Neuhauser and Raphel (2004) study large increases in benefits in California in the mid 1990's, finding higher benefit levels both increase claim frequency and also disability duration, while the additional claims filed appear to be less severe.

Logically, if moral hazard is present then one would expect relative claim rates to be higher amongst injuries or situations which exhibit greater asymmetries of information. With this theoretical prediction in mind, Bolduc et al. focus on construction workers in Ottawa, and confirm that higher benefits disproportionately increase the likelihood of filing difficult-to-diagnose injuries. Biddle and Roberts (2003) find similar evidence relating benefit generosity and claims using administrative records from Michigan, with severity of injury and overall health also playing large roles. These findings are seen as potential evidence of workers taking advantage of asymmetric information regarding the true extent of injury and recovery.

1.2.2 The Monday Effect

The first workers' compensation programs came into effect during the progress movement in the 1910's[5]; around the same time Vernon (1921) was the first to notice a Monday effect in claim rates. Decades later, Smith (1989) analyzed claims across several states, finding a disproportionate number of injuries on Monday's,

[5]See Fishback and Kantor (1995), Table II on page 722.

largest for injuries that are difficult to diagnose such as lower back injuries and sprains. In conclusion, he attributed the noticeable number of Monday claims to workers using workers' compensation to cover weekend injuries. The excess fraction of injuries occurring on Monday above 20 percent – which one would expect if work hours were distributed uniformly throughout the week – has become referred to as the "Monday Effect".

Notwithstanding those initial findings, Card and McCall (1996) and Compoliete and Hyatt (2006) offer evidence that medical insurance coverage does not influence the likelihood of a worker filing a Monday claim. Card and McCall (1996) find workers likely to have medical insurance[6] are no more or less likely to file Monday claims while Campoliete and Hyatt (2006) find a Monday effect in Canada, where public medical care is freely available. Their results suggest that employees may not be abusing workers' compensation to cover the medical costs of off-the-job injuries. However, workers could have other motives for filing off-the-job injuries through workers' compensation besides medical costs.

Because workers can replace lost wages, enjoy leisure, avoid medical deductibles, and supplement future wages with permanent disability payments, incentives for fraudulent activity remain even if workers' have medical insurance. This could be

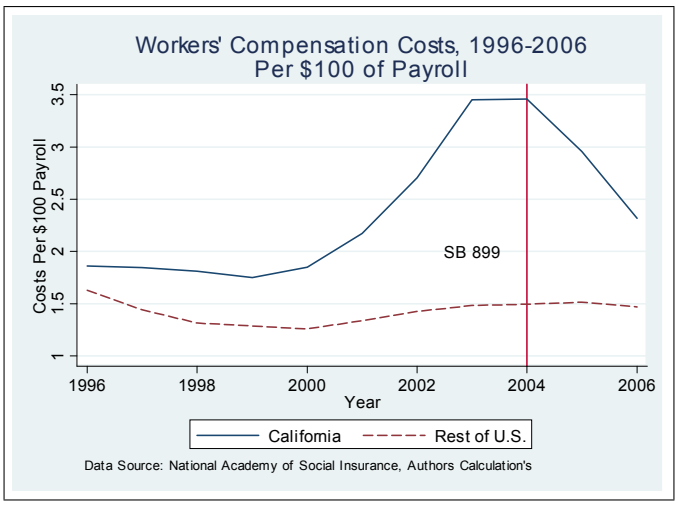
[6]Evidence from Lakdawall et al. (2007) suggests employers who offer medical insurance are also more likely to have workers' compensation claims. This could be because workplaces with large asymmetric information offer medical insurance more readily to reduce false claims, or hope that the workplace injuries may be filed through health insurance rather than workers compensation. Comparisons across medical insurance provision would need to remove such unobservables to uncover the causal effect of health care on Monday claim rates.

particularly true for injuries such as sprains and strains which require resting time in addition to medical care for recovery. In addition, Baker and Krueger (1995) and Butler et al. (1996) reveal that workers can receive greater medical coverage through workers compensation – particularly in HMO settings (because doctors receive a piece rate for treating workers' compensation injuries, and a lump sum for normal patients in an HMO). The law changes in California provide a situation to further test the Monday effect as they exogenously changed both the expected benefits and difficulties in filing false claims, while temporary employment is a situation in which pronounced asymmetric information could contain more prevalent moral hazard.

1.2.3 California Senate Bill 899

In the United States on average, insurance costs for employers fell in the early and mid 1990's. Reasons for the decline include improved workplace safety, workers' compensation reforms, and the privatization of insurance funds. Beginning in 1999 workers' compensation costs dramatically rose in California, while they slowly increased in the rest of the United States. Between 2000-2003, workers' compensation share of payroll costs nearly doubled, rising from 1.85 percent to 3.45 percent. Figure 1.1 illustrates the difference between California's workers' compensation costs as a fraction of payroll and the rest of United States.

Figure 1.1: CA vs. US Cost Trends



Due to the rising workers' compensation costs, workers' compensation was one of the focal points of recall election of 2003, and became a target for reform shortly after Governor Schwarzenegger took office.[7] With large legislative support, SB 899 was signed into law April 19, 2004 – with some provisions going into effect immediately and others on January 1, 2005. Its intent, as described by the California Division of Workers' Compensation, is to "control escalating medical costs...and compensation benefits".[8] The major reforms included allowing employers to choose the treating doctor through medical provider networks, requiring causal evidence linking the injury to the job, mandating AMA-approved objective medical standards in assessing disability, and limiting temporary total benefits to

[7] See "California Businesses Side with Schwarzenegger's Workers' Compensation Plan." Inland Valley Daily Bulletin, September 12, 2003. Also see "Davis to sign workers' comp reform bill: Issue has emerged in run-up to recall", San Diego Tribune, September 30, 2003.

[8] "Workers' compensation reforms under Senate Bill 899: First annual report of progress." California Division of Workers' Compensation

104 weeks.[9] Additional reforms included providing employers incentives to return injured workers to feasible tasks through rate reductions and requiring prompt medical care.

While requiring objective medical evidence or basing disability payments on the fraction of the injury that can be causally attributed to job tasks may seem like benign changes, they can influence the ability of a worker to a file a claim for a soft-tissue injuries such as back sprains or shoulder strains. In addition, allowing employers to choose doctors may prevent employees from finding doctors who are more willing to approve workers' compensation claims[10]. Boden and Ruser (2003) find that states who change their laws in the 1990's to requiring objective medical evidence and based disability payments on causality decreased claims. The evidence on doctor choice is mixed, as Boden and Ruser (2003) establish little evidence that medical provider networks affect claims, while Neumark et al. (2005) find that costs are higher and returning to work is delayed when workers choose their doctor. In addition, Ruser (1998) shows some evidence that employer choice of the doctor reduces the frequency of Monday Claims.

The reforms also sought to reduce injury durations. SB 899 offered employers deductions if employees were placed in different jobs with feasible tasks.[11] Tempo-

[9]See "Comission on Health and Safety and Workers' Compensation: SB 899 Topic summary report–version 4." Commission on Health and Safety and Workers' Compensation.

[10]Perhaps doctors could also end up under monopsonistic pressure from employers sending many patients to only a few doctors.

[11]Waehrer and Miller (2003) establish evidence that higher benefits and lower waiting periods increase employers' usage of restricted work.

rary total benefits[12], which before were limited to 5 years, now became restricted to 104 weeks[13]. Lastly, prompt medical care provision required employers to cover medical costs within the first 30 days[14], regardless of whether a claim is accepted or rejected.

Because of colinearity in the timing of the changes in California, rather than trying to disentangle their partial effects, this paper assesses their net effect. All of the major changes – with perhaps one exception – could conceivably make it more difficult and also less beneficial to file a on off-the-job injury as a workers' compensation claim. And while prompt medical care guarantees the initial medical coverage of all injuries, it also requires an employee reporting an injury to visit a doctor of the employer's choosing soon after the injury is reported, which could increase the likelihood that an employer-chosen doctor uncovers evidence the claim is false.

Initial evidence suggests that net effect of the reforms statewide has been achieving its goals, with costs going down and claims decreasing in number and duration. As seen in Figure 1.1, total workers' compensation costs as a fraction of payroll has fallen since the reforms, while there was no discernible change in the rest of the nation.[15] The coming section measures the size of the Monday effect

[12]Although some previously planned increases in the temporary total benefits cap went into effect in 2005 and 2006, for the temporary workers in this anlayis only 4 percent have wages which exceed the initial threshold.

[13]With a few exceptions to this included burns, eye injuries, HIV, among other severe injuries.

[14]Capped at $10,000.

[15]In addition, initial reports done by the California Workers' Compensation Institute and California Division of Workers' Compensation suggest that lost-work spell length has decreased by 17 percent.

prior to the reforms and also estimates the net effect of the reforms on Monday claims.

1.3 Initial Analysis

1.3.1 Data Source: Temporary Employment

Temporary employment has increased dramatically in recent years with its growth accounting for some 10 % of total job growth in the United States. We briefly digress to explain the labor market situation that is temporary employment, and how it relates to workers' compensation.

While temporary employment can refer to seasonal employees or outside professional consultants, we discuss workers provided by temporary agencies, which make up 71 percent of all temporary employment Dey et al. (2006). The process begins when a temporary agency recruits employees who are kept on its roster according to their skills, experience, geographic locations, and work preferences. Firms needing labor approach the temporary firm and agree to pay a wage for the employee with a mark-up to the temporary firm. The mark-up is used to cover all other costs for the workers such as payroll taxes, benefits, and workers' compensation. If the leasing firm no longer wants the employee, the employee is reassigned to positions at other firms.

So while the leasing firm controls the work environment (and therefore the

safety of the worker), the temporary agency is responsible for workers' compensation if the worker is injured. Because the temporary firm has a minimal presence at the job site, the asymmetric information between the workers and the firm is increased because monitoring workplace safety is more difficult. Furthermore the worker has minimal ties to the firm, which makes filing claims both true and false potentially less costly as workers have reduced expectations concerning promotions. Park and Butler (2001) cite these factors in explaining their empirical observation that temporary workers in Minnesota are 3-5 times more likely to file claims compared to full-time employees.[16]

Temporary employees are indeed somewhat different from the average full-time employee in the United States. Table 1.1 compares full-time employees from the February Contingent Workers Supplement for 2001 and 2005 of the Current Population Survey with a representative sample of temporary workers[17] and temporary workers from the firm under study here.[18] To summarize, temporary workers in general are younger, have less education, earn less, have lower wages, and are less likely to be married. These differences are even more pronounced for the workers from the firm in this analysis, which are in the last column. In short temporary employees on average earn less than full-time employees. Consequently, they face

[16] Another factor that could also play a role is worker inexperience.

[17] The sample is restricted to temporary workers with positive earnings age 18-65, as in the administrative data there is only earnings for temporaries who are employeed.

[18] Workers are not required to report their characteristics through the Equal Employment Opportunity Act. If workers who drop out of high school are less likely to report their education, than the statistics may understate educational and earnings differences.

greater replacement benefits because their incomes are less likely to be above maximum thresholds and more likely to qualify for the minimum payments mandated by each state.

1.3.2 Claims Data

The claims data are the population of workers' compensation claims from a large temporary staffing agency for 2002-2006. The data include both the medical-only claims and also those involving time away from work. While some previous workers' compensation research has often used claims data classified by body part or injury type, the insurance provider in this case classifies records by injury causation. The primary causes we will focus on are overexertion and blunt trauma claims. Overexertion claims are nearly always associated with some sort of soft-tissue injury (considered difficult-to-diagnose), while blunt trauma injuries involve being struck by objects typically producing fractures, lacerations, or contusions (normally considered easy-to-diagnose).[19] Summary statistics for the data are provided in Table 1.2.

1.3.3 Measuring the Monday Effect

The Monday effect has in earlier works been defined as the excess percentage of claims above 20 occurring on Monday, as that is the natural frequency that would

[19]Other easy-to-diagnose injuries such as driving accidents, burns, and eye injuries are not very common as the temporary firm seeks to avoid very dangerous jobs.

Table 1.1: Temporary vs. Full-Time Workers
Summary Statistics

	Full-Time (CPS)	Temp-Emp (CPS)	Temp-Emp (Firm)	Full-Time (CPS)	Temp-Emp (CPS)	Temp-Emp (Firm)
		Nationwide			California	
Age	38.3	37.4	31.0	37.5	37.7	31.9
Never Married	29.2	37.5	-	31.6	36.1	-
< =HS Grad	44.6	47.6	69.0	45.5	47.7	72.6
BA or More	29.0	15.2	11.2	29.5	13.8	10.0
Male	62.3	44.2	53.3	64.3	36.1	55.6
Weekly Earnings	$738.14	$507.68	$248.20	$763.87	$511.83	$253.10
Hourly Wage	$14.02	$13.01	$8.85	$14.14	$13.82	$8.96

Columns 1-2 and columns 4-5 are taken from the CPS Contingent Worker Supplements in 2001, 2005, and columns 1-2 are weighted to match with the distribution of workers from the national firm. Column 3 and 6 are from administrative and payroll records from the national staffing agency.

Table 1.2: Summary Statistics

Monday	20.6
	(40.5)
Overexertion	29.9
	(45.8)
Blunt Trauma	42.9
	(49.5)
Compensation Claim	28.2
	(45.0)
California	47.3
	(50.0)
Fraction Male	65.3
	(43.5)
Avg. Weekly Wage (Dollars)	334
	(199.5)
Avg. Weeks Worked	25.6
	(35.8)
Days Between Injury & Report	18.2
	(98.8)
Compensation	2597.3
	(10319.3)
Medical	3451.8
	(18119.6)
Legal & Travel	1402.0
	(5151.9)
Weekday Injuries	8047
Weekend Injuries	983

This table contains summary statistics for claims data in the analysis.

arise if the work hours were evenly distributed throughout the week. Table 1.3 contains a comparison of the Monday effect between the claims in this analysis and Card and McCall (1996) and Campoliete and Hyatt (2006) for difficult-to-diagnose injuries (overexertion injuries for the temporary injury claims and back injuries for Card and McCall 1996 and Campoliete and Hyatt 2006). Interestingly, the Monday effect is much stronger for compensation claims than for those claiming only medical benefits. The compensation claims (which are most comparable to the claims data in Card and McCall 1996 and Campoliete and Hyatt 2006 because their claims has only lost-workday cases) report an excess of 7 percentage points, slightly large than excess of 5 and 6 percentage Card and McCall (1996) and Campoliete and Hyatt (2006) respectively find.

That measure of the Monday effect is mainly valid if the distribution of work hours is distributed uniformly throughout the week (all injuries and types are less likely on Friday, which suggests this may not hold). Because scaling by daily hours worked is not possible, I scale by the frequency of injuries which are most likely represent the distribution of work hours because they are easy to diagnose and require immediate attention – cuts and lacerations for Card and McCall (1996) and Campoliete and Hyatt (2006) and blunt trauma injuries for the claims from the temporary firm. When scaling by the fraction of easy-to-diagnose injuries, the Monday effect falls to around 4 percentage points for both Card and McCall (1996) and Campoliete and Hyatt (2006) and increases to over

Table 1.3: Measuring Excess Monday Claims

Lost Time Back	Overexertion Injuries			Lost Time Back	
	All	Compensation	Medical	C & M	C & H
Monday Effect Above 20 Percent					
"Excess" Monday Claims	3.8***	6.8***	1.6	5.3***	6.2***
(in CA Prior to Reform)	(1.3)	(2.0)	(1.7)	(0.5)	(0.7)
"Excess" Monday Claims	3.0***	7.2***	0.9		
(outside of CA)	(1.2)	(2.0)	(1.4)		
Monday Effect Relative to Difficult-to-Diagnose Injuries					
"Excess" Monday Claims	4.4***	8.6***	1.8	4.1***	4.2**
(in CA Prior to Reform)	(1.6)	(2.7)	(2.0)	(1.3)	(2.2)
"Excess" Monday Claims	4.6***	8.3***	2.5		
(outside of CA)	(1.5)	(2.9)	(1.7)		

Notes: This table summarizes the excess fraction of claims on Mondays. The first three columns indicate the excess fraction of Monday claims for overexertion injuries by claim type for both California and other states. The forth and fifth columns are for low-back injuries from Card & McCall (1995, 1996) and Compoliete & Hyatt (2006). The first two rows report the excess fraction of difficult-to-diagnose injuries on Monday above 20 percent, while the bottom two rows indicate the excess fraction above the fraction of easy-to-diagnose injuries reported on Monday.
* sig. at 10 percent. ** sig. at 5 percent. *** sig. at 1 percent.

8 percentage points for compensation claims from the temporary firm. If workers are seeking temporary benefits to replace lost wages, the much larger Monday effect observed in the temporary firm could be explained by higher replacement rates[20] and higher degrees of asymmetric information at the job site for temporary workers[21].

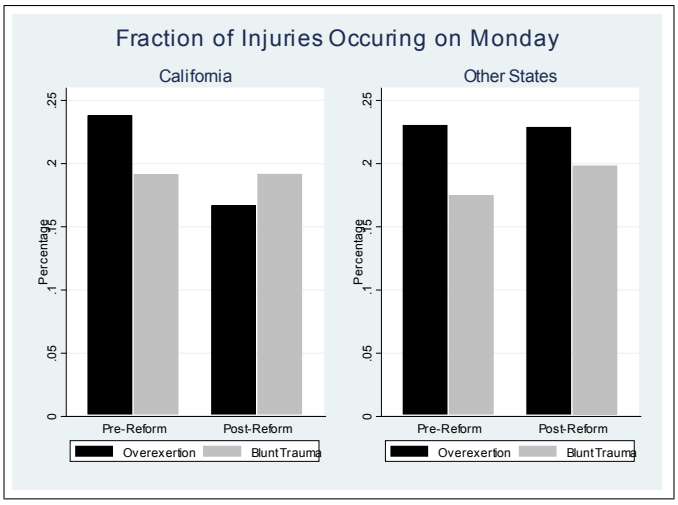
The recent reforms in California provide an exogenous shock by both increasing the difficulty of filing a false claim and also reducing the potential benefits even if such a claim is approved. Figure 1.2 presents the fraction of injuries occurring

[20]The earnings of temporary workers are less often subject to the maximum thresholds and more often subject to minimum thresholds.

[21]Another driving factor could be that few temporary workers have medical insurance, but as according to the previous, there is little evidence that medical insurance has a substantial effect on workers compensation claims.

on Mondays for compensation claims in California, both before and after the reform. Prior to SB 899, nearly 24 percent of overexertion compensation claims were reported to have occurred on Mondays in California, with only 19 percent of injuries falling on Monday for blunt trauma injuries. After the reform the fraction of Monday claims falls for overexertion injuries in California, with essentially no change for overexertion injuries outside of California or blunt trauma injuries inside of California, to comparison groups that might indicate whether there was a substantial shift in work hour distribution.

Figure 1.2: Monday Effect Beform and After Reforms



The before-after comparison of effects of the laws is explored in more detail in Table 1.4. It contains the before after comparisons by the cause of injury (overexertion, blunt trauma), and location (in California, out of California). After the law changes (claims in 2005-2006[22]), California shows significant changes for injuries

[22]Because the some of the intital reforms went into effect in 2004, with the rest (AMA guide-

Table 1.4: Before and After Comparisons of the Monday Effect

Location	Monday Injuries Relative Frequency		Difference	|T-Test|	Diff-in-Diff	|T-Test|
	Before-After Comparisons					
	2002-2004	2005-2006				
	(1)	(2)	(3)	(4)	(5)	(6)
All Injuries						
CA	20.8	19.6	-1.1	0.78	-1.7	0.89
Not-CA	20.6	21.1	0.6	0.46	-	-
Overexertion						
CA	23.9	16.7	-7.2	2.64	-6.9	1.84
Not-CA	23.1	22.8	-0.3	0.14	-	-
Blunt Trauma						
CA	19.1	19.0	-0.1	0.10	-3.2	1.08
Not-CA	17.2	20.2	3.0	0.54	-	-

Notes: This table shows the Fraction of Injuries Occurring on Monday, both before and after the law change.

Statistically significant results are highlighted in bold.

* sig. at 10 percent. ** sig. at 5 percent. *** sig. at 1 percent.

whose cause is overexertion. The fraction of claims on Mondays for overexertion injuries falls by 7.2 percentage points. Adjusting for Monday claiming frequencies in other states, this changes only slightly to 6.9. The other injuries or claim types experience no significant changes in a statistical sense, and most are small in magnitude as well. The same can be said for all injury and claim types occurring in branches outside of California, showing no significant reductions in relative Monday injury rates.

lines and doctor choice) goign into effect on Jan. 1, 2005. I have tried models both excluding the data from 5/2004-12/2004, or creating sepearte indicators to parse out those effects, and find no distinguishable differences.

1.4 Monday Claims: Regression Results

The initial evidence suggests that during the post-reform period the number of Monday injuries for difficult-to-diagnose claims fell in California. Further analysis in regressions allows one to control for occupation and individual characteristics. We proceed with linear probability models where the dependent variable is an indicator for whether or not the injury occurred on a Monday. taking the form of equation (1.1). The regressions control for the occupation (taken from the workers' compensation code), sex, state, replacement rate, insurance rate and are clustered by state (standard errors in difference-in-difference models use block bootstraps[23]). The main coefficient of interest is the indicator for whether the injury occurred in the post-reform period, 2005-2006. The final column takes the form of equation (1.2), where the effect of the policy will be measured by the interaction between a California indicator and an indicator for the post-reform period.

$$Monday_{iost} = \beta_o + X_i'\alpha + S_s + \delta * after\_reform_t + u_{iost} \tag{1.1}$$

$$Monday_{iost} = \beta_o + X_i'\alpha + S_s + \delta * after\_reform_t + \gamma * CA * after\_reform_t + u_{iost} \tag{1.2}$$

[23]See Bertrand et al. (2004) and Cameron et al. (2007).

In each of the regressions i is a claim, β_o is an occupation fixed effect, X_i' is the vector of controls for the individual claim, S_s is a state fixed effect, CA is an indicator for CA, $after\_reform_t$ is an indicator for if the injury occurred after all of the reforms were in place. If the distribution of work hours remained constant in California, then specification (1.1) is sufficient to measure the decrease in the relative frequency of Monday injuries. If there were unobserved changes in the distribution of work hours that are similar within occupations and across states, specification (1.2) can adjust for such shifts. On the other hand, estimating the specifications for easy-to-diagnose injuries offers an additional robustness test if there were a change in work hours specific to California, as there would be a corresponding change in the probability of a Monday easy-to-diagnose injury.

The results in Table 1.4 confirm the previous summary statistics and suggest that the fraction of Monday injuries decreased in California for difficult-to-diagnose claims. For most of the specifications chosen, there appears to be no effect on Monday claims for more easy-to-diagnose blunt trauma injuries. In addition, the estimates of the decrease are quite similar across the two specifications at -0.076 for the first difference specification and -0.069 for the difference-in-difference model. With this in mind, the estimates suggest that the net of the California reforms might have eliminated the excess fraction of Monday claims for difficult to diagnose injuries.

Table 1.5: Reform Impact on the Probability of Monday Claim

Injury	Claim Type	First Difference		Diff-In-Diff	Diff-in-Diff
		CA	Not-CA		
		(1)	(2)	(3)	(4)
All	All	-0.021	0.0048	-0.020	0.026
		(0.016)	(0.013)	(0.014)	(0.022)
Overexertion	All	-0.076**	0.0067	-0.069*	-0.084
		(0.030)	(0.027)	(0.037)	(0.050)
Blunt Trauma	All	-0.004	0.021	-0.023	0.014
		(0.028)	(0.020)	(0.022)	(0.038)
Controls					
State FE		N/A	Yes	Yes	Yes
State Linear Trends		No	No	No	Yes

Notes: The dependent variable in all regressions is whether a claim occurred on a Monday, estimated by linear probability models. Included controls are state and occupation fixed effects, weeks worked, sex, insurance rate, and wage replacement rate. First difference models use heteroskedastic robust standard errors while the difference-in-difference models cluster by state and use block bootstrap errors.

* sig. at 10 percent. ** sig. at 5 percent. *** sig. at 1 percent.

Earlier it was shown that the Monday effect was largest for difficult-to-diagnose injuries seeking compensation benefits, and there was only a slight excess fraction of Monday claims for medical only causes. In Table 1.6 the same regressions as Table 1.5 are estimated, separating the results by whether the claim was for compensation, or only related to medical expenses. Just as the Monday effect was largest for overexertion injuries claiming compensation earlier, that same group has the numerically largest decrease in the probability Monday injuries following the reforms. The first difference model estimate the reduction in the probability of Monday claims to be -0.09, estimated to be -0.13 for the difference-in-difference specification. Both are within the neighborhood of the 0.086 excess probability of Monday injury for overexertion compensation claims (although the difference-in-difference estimator is slightly more noisy). Once again, the probability of a Monday injury for blunt trauma claims is minimally affected by the reforms.

Table 1.6: Reform Impact on Monday Effect, by Claim Type

Claim Type	Compensation		Med-Only	
	(1)	(2)	(3)	(4)
Overexertion	-0.09**	-0.13*	-0.06	-0.03
	(0.04)	(0.07)	(0.05)	(0.02)
Blunt Trauma	0.02	0.03	-0.02	-0.04
	(0.05)	(0.06)	(0.04)	(0.04)
"Monday Effect" Prior to Reforms	0.09		0.02	
Specification	First Diff	Diff-in-Diff	First Diff	Diff-in-Diff

Notes: The dependent variable in all regressions is whether a claim occurred on a Monday, estimated by linear probability models. Included controls are state and occupation fixed, sex, insurance rates, and the wage replacement rate. First difference models are CA only and use heteroskedastic robust standard errors. Difference-in-difference models use block bootstrap errors and cluster by state.

* sig. at 10 percent. ** sig. at 5 percent, *** sig at 1 percent.

Both the simple summary statistics in Section 3 and the regression results here suggest similar conclusions. Following the reforms in California, the fraction of overexertion occurring on Monday fell. Furthermore, this decrease is largest and most statistically significant for compensation claims rather than medical only injuries. The decrease in the probability of a Monday injury following the reforms is consistent with a model where the Monday effect – or some fraction of it – is due to off-the-job injuries.

1.5 Claim Rates and Costs

1.5.1 Costs Per Claim

While up to now this paper has found evidence that the reforms in California influenced the probability of a worker filing a Monday claim, the main purpose of the SB 899 was to reduce claim costs, or from the workers' standpoint, lower benefits. For the temporary firm in the pre-reform period, compensation, medical and legal costs were respectively 114, 50, and 102 percent higher in California than costs per claim in other states. Regression models of the form of equation (1.3) estimate the percentage effect of the reform inside and outside of California, while those of equation (1.4) in the later model estimate the relative change in California. These regressions adjust for the same of controls as linear probability models in Section 4. As in Butler et al. 1997, incurred costs are available due to the use of administrative micro-claim data.[24] The effect of the reforms on compensation, medical, and legal costs is assessed in Table 1.7.

$$expense_{iost} = \beta_o + X_i'\alpha + S_s + \delta * after\_reform_t + u_{iost} \tag{1.3}$$

[24] 98 percent of the claims are closed, suggest the costs per claim are relatively complete.

$$expense_{iost} = \beta_o + X_i'\alpha + S_s + \delta * after\_reform_t + \gamma * CA * after\_reform_t + u_{iost} \tag{1.4}$$

In the post-reform period, costs fell in California both relative to previous claims and adjust for changes in for temporary workers in other states. The compensation costs for all claims is estimated to fall by 48 percent relative to claims in other states, while medical and legal costs fall by 56 and 40 percent.[25] While each of these decreases is substantial, recall the margin by which California costs exceeded those from other states in the pre-reform period. While the reforms have decreased the gap between California and claims from other states, in the post-reform period costs per claim continue to be somewhat higher for the compensation, medical, and legal categories – respectively by 45, 1, and 38 percent.

[25]This scaling the dollar decrease in costs by the 2002-2004 average costs for California which were respectively $4,561, $5,142, and $2,451 for compensation, medical, and legal/travel expenses paid.

Table 1.7: Reform Impact on Claim Costs/Benefits

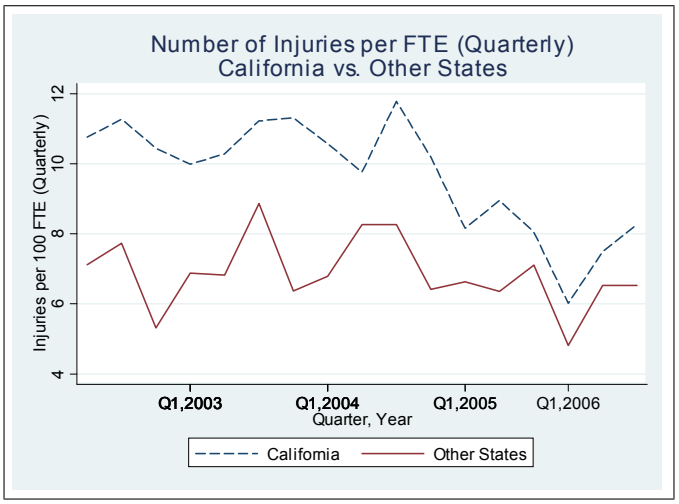
Expense	Injury	First Difference		Diff-in-Diff
		California	Not-California	
Compensation	All	-2627.32***	-221.36	-2178.79**
		(758.34)	(359.90)	(1056.50)
Compensation	Overexertion	-2578.51***	111.32	-2247.88**
		(464.5465)	(246.52)	(1078.54)
Compensation	Blunt Trauma	-2637.63***	128.54	-2010.50**
		(830.88)	(313.36)	(1005.07)
Medical	All	-3224.15***	-115.14	-2902.02**
		(535.70)	(464.77)	(1395.11)
Medical	Overexertion	-4093.28***	714.87**	-4252.30**
		(844.94)	(241.52)	(1993.27)
Medical	Blunt Trauma	-3296.69***	-1178.41	-2317.47*
		(1088.73)	(1251.10)	(1300.6)
Legal	All	-1217.30***	26.96	-981.95**
		(227.15)	(107.45)	(484.85)
Legal	Overexertion	-1429.05***	51.20	-1192.85**
		(414.14)	(168.18)	(592.91)
Legal	Blunt Trauma	-957.14**	32.673	-562.91
		(422.25)	(170.04)	(343.31)

Notes: The dependent variable in all regressions is the benefits/costs for a claim estimated by OLS. Controls are state and occupation fixed effects, sex, insurance rate, and wage replacement rate. The first difference models use heteroskedastic robust standard errors, while difference-in-difference models use block bootstrap errors and cluster by state.
* sig. at 10 percent. ** sig. at 5 percent, *** sig at 1.

1.5.2 Claim Rates

In addition to affecting the relative frequency of Monday claims, the reforms in California could also reduce the aggregate frequency of injuries. Due to the decreases in potential benefits caused by more objective standards in assessing disability and also limits in temporary disability payment length, workers might see less returns in filing claims for injuries. Furthermore they could also exhibit more effort to reduce their exposure to danger at work. Figure 1.3 presents the number of claims per FTE filed over 2002-2006 both inside and outside of California. By the first quarter of 2005, the number of claims falls in California but experiences only a slight decline in branches in other states. This is similar to evidence from Butler et al. 1997, who also study a single large employer and find that claim rates decrease with lower benefit levels.

Figure 1.3: Injury Rates 2002-2006



The magnitude of the decrease in claim rates is explored in further detail in Table 8. In order to estimate the effect of the reforms on claim rates, once again claims in other states form a counter-factual group. The injuries are normalized by either full-time equivalents (FTE) in equation (1.5), or alternatively by workers compensation premiums (WCP) paid in equation (1.6).[26] The log of the injury rate with either normalization is the dependent variable in the regression models, which allows the coefficients to be interpreted as percentage effects. A monthly time series is constructed for both California and all offices outside of California over the 2002-2006 sample, and month dummy variables are included to control for seasonality.

$$\ln\left(\frac{injuries}{FTE}\right)_{mct} = m_m + after_t + CA + \gamma * CA * after_t + u_{mct} \tag{1.5}$$

$$\ln\left(\frac{injuries}{WCP}\right)_{mct} = m_m + after_t + CA + \gamma * CA * after_t + u_{mct} \tag{1.6}$$

[26]Previous studies have often relied only claim rates adjusted by FTE, and constructing injury rates by industry or occupation codes to absorb differences in risks associated with industry or occupation shifts. This is attractive in settings with enough claims to avoid many occupations and industries having 0 injuries. While the individual workers compensation occupation codes are available for employees and injured workers in the administrative data, aggregating to only workers' compensation occupation code level could introduce many zero counts. By instead aggregating total hours weight occupation cells by 2006 premiums (to keep the measure of risk associated constant), produces a measure of weighted employment that is essentially FTE weighted by relative risk. Thus a monthly time series is constructed for both California and all offices outside of California over the 2002-2006 sample, and month dummy variables are included to control for seasonality.

Table 1.8: Reform Impact on Claim Rates

	Normalizing Factor	First Difference		Diff-in-Diff
		California	Not-California	
		(1)	(2)	(3)
All Injuries	FTE	-0.30***	-0.12*	-0.19**
		(0.06)	(.06)	(0.07)
Overexertion	FTE	-0.40***	-0.11	-0.34**
		(0.10)	(0.09)	(0.13)
Blunt Trauma	FTE	-0.56***	-0.10	-0.44***
		(0.09)	(0.07)	(0.11)
All Injuries	WCP	-0.12*	0.076	-0.16**
		(0.06)	(0.077)	(0.08)
Overexertion	WCP	-0.22**	0.07	-0.32**
		(0.10)	(0.10)	(0.13)
Blunt Trauma	WCP	-0.37***	0.08	-0.42***
		(0.09)	(0.08)	(0.11)

Notes: These regressions use an aggregated monthly time series of the log of the number of injuries normalized by FTE or the workers' compensation insurance paid. All OLS regressions include monthly indicators to adjust for seasonality, and regressions report robust standard errors which were 10 percent larger than those correcting for first-order autocorrelation.

* sig. at 10 percent. ** sig.. at 5 percent. *** sig. at 1 percent.

The point estimates from column 3 (which adjust for common trends to the company or nation) of Table 1.8 suggest the total claims per FTE or WCP fell by either 19 or 16 percent, respectively. Similarly, total overexertion claims decreased by 34 or 32 percent, while blunt trauma injuries declined by 44 or 42 percent. The difference between normalizing factors in column 3 are minimal, with point estimates remaining robust to either normalizing factor.[27]

[27]Another approach would be to run a linear probability model or probit for whether or not an injury occured for using each individual worker. Results for each individual workers from 2002-2006 are reported in Table 3 in the appendix. Table 1.4 in the appendix runs similar models to equation 1.6 breaking down the difference by medical only and compensation injuries.

Table 1.9: Hypothetical Counterfactual Costs 2002-2004

	2002-2004	Counter-factual
Claims	3137	2541
Per Claim Cost	$12,161	$6,098
Total Cost	$38,150,000	$15,500,000
Excess Overexertion Monday Claims/Costs		
Overexertion Claims	65	0
Total Cost of Monday Effect (02-04 costs)	$790,000	0
Total Cost of Monday Effect (05-06 costs)	$396,000	0
Monday Effect's	3.5	1.8
Percentage of Cost Reduction	(cost 02-04)	(cost 05-06)

Notes: This table presents a counter-factual view of what costs and claims would have been if the reforms had already been in effect during 2002-2004. The last row shows what fraction of the reduction in total costs can be attributed to the reduction in Monday difficult-to-diagnose injuries. The calculations are based on the estimates from Tables 1.5, 1.7, and 1.8.

1.5.3 Overall Cost Analysis

Through the analysis in this paper, it has been shown that the fraction of injuries occurring on Monday for difficult-to-diagnose fell in California, as did claim rates and claim costs. This section uses the previous estimates to provide a relatively simple counter-factual view of the world. If the reforms had gone into effect earlier in California, what would have been the claim rates, costs, and the excess number of Monday injuries. The results from Tables 1.5, 1.7 and 1.8 are used in the construction of the counter-factuals and are presented in Table 1.9.

As shown in Table 1.9, the total costs due to injuries have fallen substantially since the reforms. Combining the reduction in benefits with the decrease in overall claims rates, the overall costs of injuries for 2002-2005 would have been nearly 23

million dollars lower if the reforms had already been in place. Of this, the reduction of the Monday effect for difficult-to-diagnose injuries accounts for between 1.8 to 3.5 percent of the total decrease in costs, and at most 8 percent[28] of the decrease in claims.

1.6 Conclusions

This paper provides evidence on the excess number of Monday injuries in workers' compensation. Using detailed claims and employment data from a large temporary agency, it is shown that major reforms in California due to SB 899 were followed by a reduction in the number of Monday injuries for difficult-to-diagnose claims. Similarly, both the number of claims per FTE and costs per claim fell absolutely and relative to branches in other states following the reforms.

Given this evidence, can one infer that the Monday effect – or some fraction of it – is due to weekend injuries being filed through workers' compensation? Consider some other physiological explanations such as weekend over-activity. Because of the decrease in the total number of injuries following the reforms, one could argue that workers are exhibiting more effort in safety at the job-site. If the increased safety effort of employees is making them less prone to injury in general, one could argue that the higher safety effort levels would also make Monday

[28]If one measures the excess number of Monday injuries as defined by compensation claims, as that was the only subcategory significant on its own, it amounts to only 5 percent.

injuries whose source is weekend activity or inactivity – and not fraudulent claims – less likely.

However, the Monday effect is largest for claims seeking compensation benefits, which suggests that the Monday effect may be driven by workers seeking time away from work in additional to medical benefits. Furthermore, the compensation-related overexertion injuries, the group with the largest Monday effect, are also those with the largest decrease in the probability of filing a Monday claim following the reforms. Given these additional findings, the evidence from this large temporary firm and the policy changes in California is most consistent with a model where some fraction of the Monday effect can be attributed to off-the-job injuries.

Lastly, it must be noted that although there is evidence that the substantial reforms in CA are associated with a reduction of excess Monday claims for difficult-to-diagnose injuries, both overall claim costs and claim rates were also affected by the policy changes. When accounting for these differences, the cost of claims filed in 2002-2004 would have been reduced by $23,000,000 in CA. Of this, the elimination of excess Monday injuries amounts to at most $630,000, or 3.5 percent of the total reduction in costs. With that in mind, although policies may exist which can reduce the excess number of Monday claims being filed, any differential effects on Monday claims will most likely be dwarfed by other first-order responses in claiming behavior.

Bibliography

[1] Baker, L. C. and A. B. Krueger (1995). "Medical Costs in Workers' Compensation Insurance. *Journal of Health Economics*, **14,** 531-549.

[2] Bertrand, M., E. Duflo, and S. Mullainathan (2004). "How Much Should We Trust Difference in Difference Estimators?" *Quarterly Journal of Economics,* **119**, 249-275.

[3] Biddle, J. and K. Roberts (2003). "Claiming Behavior in Workers' Compensation." *The Journal of Risk and Insurance*, **70**, 759-780.

[4] Boden, L. I., and J. W. Ruser (2003). "Workers' Compensation 'Reforms', Choice of Medical Care Provider, and Reported Workplace Injuries." *Review of Economics and Statistics,* **85**, 923-929.

[5] Bolduc, D. B. Fortin, F. Labreque and P. Lanoie (2002). "Workers' Compensation, Moral Hazard, and the Composition of Workplace Injuries." *The Journal of Human Resources,* **37,** 623-642.

[6] Butler, R. J., D. L. Durbin and H. H. Gardner (1996). "Increasing Claims for Soft Tissue Injuries in Workers' Compensation." *Journal of Risk and Uncertainty,* **13,** 73-87.

[7] Butler, R. J., B. D. Gardner, and H. H. Gardner (1997). "Workers' Compensation Costs When Maximum Benefits Change." *The Journal of Risk and Insurance,* **13,** 259-269.

[8] Butler, R. J. and J. D. Worrall (1983). "Workers' Compensation: Benefit and Injury Claim Rates in the 1970's. " *Review of Economics and Statistics,* **50,** 580-589.

[9] Butler, R. J. and J. D. Worrall (1985). "Work Injury Compensation and the Duration of Nonwork Spells." *Economic Journal,* **95,** 580-589.

[10] Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). "Bootstrap-based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics*, **90**, 414-427.

[11] Compliete, M. and D. E. Hyatt. "Futher Evidence on the Monday Effect in Workers' Compensation." *Industrial and Labor Relations Review*, **59**, 438-450.

[12] Card, D. and B. P. McCall (1995). "Is Workers' Compensation Covering Uninsured Medical Costs: Evidence from the 'Monday Effect'." *NBER Working Paper No. 5058.*

[13] Card, D. and B. P. McCall (1996). "Is Workers' Compensation Covering Uninsured Medical Costs: Evidence from the 'Monday Effect'." *Industrial and Labor Relations Review,* **59**, 438-450.

[14] Card, D. and B. P. McCall (1995). "When to Start a Fight and When to Fight Back: Liability Disputes in the Workers' Compensation System." *NBER Working Paper No. 11918.*

[15] Dey, M., S. Houseman, and A. Polivka (2006). "Manufacturers Outsourcing to Employment Services." *Upjohn Staff Working Paper No. 07-13.*

[16] Fisback, P. V. and S. E. Kantor (1995). "Did Workers Pay for the Passage of Workers' Compensation Laws." *Quaterly Journal of Economics,* **110,** 713-742.

[17] Fortin, B. and P. Lanoie (1992). "Substitution Between Unemployment Insurance and Workers' Compensation: An Analysis Applied to the Risk of Workplace Injuries." *Journal of Public Economics,* **49**, 287-312.

[18] Krueger, A. B. (1990). "Incentive Effects of Workers' Compensation." *Journal of Public Economics,* **41**, 73-99.

[19] Lakdawalla, D. N., R. T. Reville and S. A. Seabury. "How Does Health Insurance Affect Workers' Compensation Filing." *Economic Inquiry,* **45,** 286-303.

[20] Neuhauser, F. and S. Raphel (2004). "The Effect of an Increase in Workers' Compensation Benefits on the Duration and Frequency of Benefit Receipt." *Review of Economics and Statistics,* **86**, 288-302.

[21] Neumark, D., P. S. Barth and R. Victor (2005). "The Impact of Provider Choice on Workers' Compensation Costs and Outcomes." *NBER Working Paper No. 11855.*

[22] Park. Y. S. and R. J. Butler (2001). "The Safety Costs of Contingent Work: Evidence from Minnesota." *Journal of Labor Research,* **65**, 101-124.

[23] Ruser, J. W. (1998). "Does Workers' Compensation Encourage Hard to Diagnose Injuries." *Journal of Risk and Insurance,* **65,**101-124.

[24] Smith, R. S. (1989). "Mostly on Monday: Is Workers' Compensation Covering Off-the-Job Injuries." In *Benefits, Costs and Cycles in Workers' Compensation.* Norwood, Mass: Kluwer Academic Publishers.

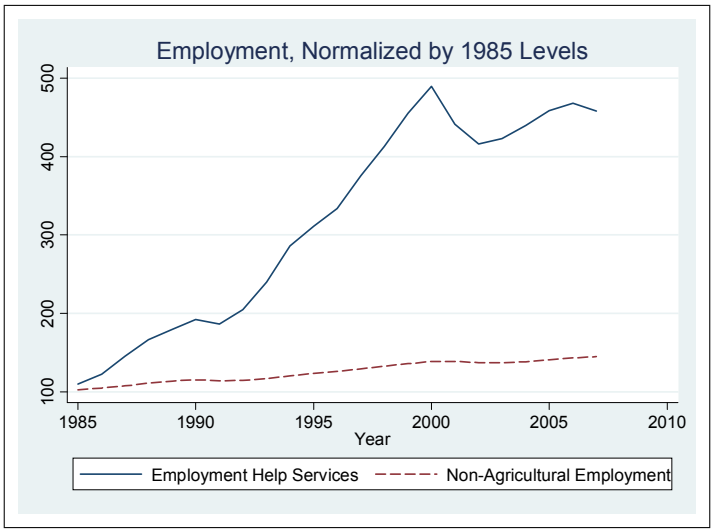
[25] Vernon, H. M. (1977). *Industrial Relations and Fatigue.* New York, Arno Press (Orignally London: George Routledge and Sons in 1921).

[26] Virtanen, M., M. Kivimki, M. Joensuu, P. Virtanen, M. Elovainio and J. Vahtera (2005). "Temporary Employment and Health: a Review." *International Journal of Epidemiology,* **34,** 610-622.

[27] Waehrer, G. M. and T. R. Miller (2003). "Restricted Work, Workers' Compensation and Days Away from Work." *Journal of Human Resources,* **38,** 964-991.

1.7 Appendix

Appendix Figure 1.1



Data Source: Current Employment Survey, Author's calculations

Appendix Table 1.1: Employment/FTE By State

State	Employees	FTE
AL	334	36.2
AR	1818	234.9
AZ	9928	1771.0
CA	112689	35659.8
CO	8726	1575.0
CT	1066	295.8
DE	1233	209.8
FL	28941	4876.9
GA	6394	1295.6
HI	1909	331.6
IA	8260	1728.3
ID	1409	161.2
IL	15288	2445.9
IN	2717	421.6
KS	548	99.4
KY	5653	731.9
LA	6800	1331.3
MA	3414	756.5
MD	1705	226.7
MI	2640	472.9
MN	312	70.2
MO	2405	322.9
MS	629	5.70
NC	4034	663.9
NE	4324	589.8
NH	189	33.1
NJ	2389	440.8
NM	50	0.07
NV	3994	584.4
NY	5017	767.3
OH	28647	4215.3
OK	6303	939.9
OR	2376	292.8

Employment/FTE By State

Appendix Table 1.1 (cont): Employment/FTE By State

State	Employees	FTE
PA	7832	1857.3
SC	4150	938.1
TN	24587	6080.1
TX	32726	6839.0
UT	1949	368.0
VI	10688	1777.5
WA	1921	5244.4
WI	9716	1418.6

Employment/FTE By State

1.7.1 Litigation

While workers compensation is intended to be a no-fault insurance system, many claims still end up in litigation. This could be because firms believe the claim to be false, or the firm could act strategically to deny claims they believe will "go away" Card and McCall (2006). For the firm in question, roughly 10 percent of all claims are litigated, with nearly 30 percent of compensation claims resulting in legal dispute. The model in Card and McCall (1995) suggests that employers are more likely to litigate claims they believe to be false. If a disproportionate number of Monday injuries were due to fraudulent claims, firms would have incentives to more closely monitor such claims. However, Card and McCall (1996) find that Monday claims were no more or less likely to be denied than claims on other days of the week.

An analysis of the compensation claims in Appendix Table 1.2 reveals that for the temporary firm in this analysis, Monday claims appear more likely to be litigated outside of California. In addition outside of California, the odds of litigation increase with the delay between the reporting of a claim is delayed and its reported date of occurrence. In California, the day of week does not strongly effect the odds litigation for claims (this true for both the pre and post-reform period, while they are reported together). However a detailed report by the Rand Institute for Civil Justice in 2003 found that litigation in California often occurs even where there are no disputes, and along with this finding delaying the filing

of a claim has no bearing on litigation in California.

Appendix Table 1.2: Probability of Litigation

	Outside of California			California		
Dependent Variable	All Injuries	Overexertion	Blunt Trauma	All Injuries	Overexertion	Blunt Trauma
All Claims						
Monday	−0.0028	.047***	−0.031	0.008	0.021	0.021
	(0.0073)	(0.017)	(0.0075)	(0.014)	(.027)	(0.027)
Days (100's) from Inj. Rpt	0.014	0.041	−0.005	0.0024	0.0014	−0.0039
	(.010)	(0.027)	(0.014)	(0.0039)	(.0062)	(0.021)
Compensation Claims						
Monday	−0.029	0.085*	−0.14***	−.014	−0.045	0.012
	(.019)	(.045)	(0.037)	(.040)	(0.056)	(0.074)
Days (100's) from Inj. Rpt	0.18***	0.24***	0.07	−0.0145	−0.0145	−0.02
	(.054)	(0.090)	(0.06)	(0.006)	(0.008)	(0.017)
State Controls	*Yes*	*Yes*	*Yes*	*N/A*	*N/A*	*N/A*
Occupation Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Individual Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes: The dependent variable in the regressions is whether or not the claim is litigated estimated by linear probability models. Controls include state and occupation fixed effects, the replacement rate, sex, weeks worked, and time between injury and filing. All regression use robust standard errors. * sig. at 10 percent. ** at 5 percent, *** sig at 1 percent .

1.7.2 Additional Claim Results

Appendix Table 1.3 contains a linear probability model which estimates whether or not an injury occurred for each employee from 2002-2006. Controls are included similar to the previous regressions, also controlling for insurance risk associated with the occupation and also whether the worker had a criminal history. We find very similar estimates to Tables 1.5 and 1.8 – with both the incidence of injuries has going down in California following the reforms, and a reduction in Monday injuries that is specific to difficult-to-diagnose causes.

Appendix Table 1.3: Probability of Injury(*100) using All Workers

	Injured	Overexertion	Blunt Trauma	Monday Overexertion	Monday Blunt Trauma
CA*After	-0.27***	-0.1**	-0.26***	-0.078***	-0.003***
	(0.1)	(0.05)	(0.1)	(0.009)	(0.002)
Ref. Assoc.	0.24***	0.06**	0.12***	0.02*	-0.016
	(0.06)	(0.03)	(0.04)	(0.1)	(0.01)
Criminal	0.39*	-0.04	0.36*	-0.013	0.15*
	(0.2)	(0.2)	(0.15)	(0.04)	(0.07)
Avg. Gross	-0.01***	-0.004***	-0.004***	-0.0009***	-0.0004
	(0.001)	(0.0001)	(0.0001)	(0.0002)	(0.0005)
Avg. Hours	0.3***	0.06***	0.01***	0.002***	0.001*
	(0.02)	(0.009)	(0.005)	(0.0004)	(0.006)
Weeks Worked	0.012***	0.002***	0.012***	0.0003**	0.001***
	(0.001)	(0.001)	(0.001)	(0.0001)	(0.0003)
Mean Injury Rate in CA Prior to Reforms	2.0	0.6	1.00	0.13	0.12

Notes: The regressions estimate linear probability models for whether or not an injury is reported by a worker. Controls include gross earnings, sex, weeks worked avg. hours weekly hours weekend, criminal conviction, whether they were referred by another employee, occupation fixed effects, state fixed effect, insurance rates. All regression report robust standard errors. All estimated coefficients are multiplied by 100. * sig. at 10 percent. ** sig. at 5 percent. *** sig. at 1 percent.

Appendix Table 1.4 contains the difference-and-difference claim per fte results separated by medical only and compensation claims. Columns 1 and 2 of

Appendix Table 1.4 respectively report the number of medical only versus compensation claims. Given that medical only claims are also falling, this could be evidence that workers are being safer at the job site. The effect on compensation claims is similar in magnitude for overexertion and blunt trauma claims – with the estimates for overexertion exhibiting slightly more noise. The effect on blunt trauma medical claims is much larger than the medical effect for overexertion claims, suggesting that the decrease in claims amongst the blunt trauma injuries could be partially driven by increased safety effort amongst employees.

Appendix Table 1.4: Effect of Reforms on Claim Rates, by Claim Type

	Normalizing Factor	Medical	Compensation
		(1)	(2)
All Injuries	FTE	-0.19**	-0.25*
		(0.09)	(0.13)
Overexertion	FTE	-0.39**	-0.30
		(0.18)	(0.18)
Blunt Trauma	FTE	-0.51***	-0.29**
		(0.14)	(0.15)
All Injuries	WCP	-0.16*	-0.23*
		(0.09)	(0.13)
Overexertion	WCP	-0.37**	-0.28
		(0.17)	(0.19)
Blunt Trauma	WCP	-0.49***	-0.27*
		(0.14)	(0.16)

Notes: These regressions use an aggregated monthly time series of the log of the number of injuries normalized by FTE or the workers' compensation insurance paid. These regressions are difference-in-difference models and are comparable to equation 6. All regressions include monthly indicators to adjust for seasonality, and regressions report robust standard errors which were 10 percent larger than those correcting for first-order autocorrelation.

* sig. at 10 percent. ** sig. at 5 percent. *** sig. at 1 percent.

1.7.3 Robustness Checks: Placebo Treatments 1998-2001

In econometric analysis of policy changes using difference-in-difference style estimation, it is important to consider the role autocorrelation can play (Bertrand et. al 2004 and Cameron et. al 2007). As shown in Bertrand et. al (2004), failing to account for such dependence in the error terms can lead to over-rejection of the null hypothesis. There are several reasons this issue would be less severe in our data. First, the sample time period under consideration is relatively short. In addition, while wages, a variable that has consistently growing on average over time–are studied in Bernard et.al, the reasons for the fraction of claims occurring on Monday exhibiting strong dependence is not obvious. Nonetheless, additional claims data for the same company from 1998-2001 provide a potential placebo test group. To test what effects autocorrelation could play in detecting changes in the relative frequency of Monday claims or claim costs we randomly generate laws for a time period under which there is no large changes in laws. We find no evidence of over-rejection in Appendix Table 1.5 suggesting that the size in our tests may be close to the nominal level.

Appendix Table 1.5: Null Rejection Frequency for Placebo Treatment Groups

	All Claims	Medical	Compensation
Monday Claims			
	0.050	0.032	0.054
Claim Costs			
	0.048	0.05	0.044

Notes: This table presents null rejection frequency when treated and control groups were randomly assigned during the 1998-2001 years, a period where no substantial policies changes took place. Using the same estimation and clustering strategy the rejection were at levels close to the nominal level.

Chapter 2

School Year Length and Student Performance:

Quasi-Experimental Evidence

2.1 Introduction

The positive association between education and earnings is one of the most robust findings in labor economics. However, not all educations are created equal. Indeed, quality has varied historically across demographic groups both within the United States and across countries. As a result, some policy makers have suggested increasing the quality of education as a tool in reducing labor market gaps in wages and employment. Interestingly, the pupil teacher ratio and per student spending – common policy interventions – have respectively fallen and risen in recent years while school year length has been stable (see Figure 2.1). Longer school years provide the potential for increased instruction time, review, and attention for individual students. If increased school year length does improve student performance, it could also be an alternative input for schools. This paper offers quasi-experimental evidence concerning instructional days and and student performance.

There is a continuing debate on whether educational quality has a bearing on student outcomes – with academics, educators, and policy makers on both sides. The discourse began with the Coleman Report (1966), which found that per pupil resources have little impact on student success. Since then, for every study refuting the Coleman Report's conclusions, another supports them. Hanushek (1981) shows increased expenditure on teachers is unlikely to improve performance. Meanwhile, Margo (1986) estimates that 27 percent of the black/white

literacy gap from 1920 to 1950 can be explained by differences in educational inputs. Krueger (1999) finds Project STAR students randomly assigned to small classes do better on standardized exams, though the benefits may be temporary. The overall consensus has been a lack thereof. Relatively little work has investigated the impact of school year length, but that done has continued in the same spirit of discord.

Initial research on school year length focused on labor market outcomes, while later studies have investigated test scores. Card and Krueger (1992) compare workers raised in different states, finding those from states with relatively longer school years earn more. Pischke (2003) takes advantage of short school years mandated in Germany to unify their schooling system.[1] He concludes that shorter school years increase grade repetition, but have no long-term effects on employment or wages. Contrarily, recent international cross-section studies by Lee and Barro (2001) and Wobmann (2000) conclude school year length has no impact on test scores. Eren and Mittlemet (2005) study the National Longitudinal Survey of Youth, which asks whether an institution's school year is longer or shorter than 180 days. They find that the best performing students benefit from longer school years while low performing students do worse with increased instructional time. Marcotte (2007) investigates the reduced form relationship between yearly snowfall and test scores, finding years with substantial snowfall are associated

[1]Through a similar regime change, Krashinsky (2006) studies the elimination of the fifth year of high school in Ontario, Canada. Cohorts with four years of high school had substantially lower grade point averages in college than those who attended high school for five years.

with lower performance in Maryland. Like previous school quality research, a consensus has yet to be reached regarding school year length's effect on student outcomes.

Due to inclement weather, districts routinely cancel school to avoid the liability and danger of traveling on unsafe roads. These cancellations, commonly called "snow days", vary from year to year and across districts, causing states to adopt policies in order to guarantee school is in session sufficiently. For example, the state of Colorado mandates that schools must extend their school year into the summer if total instructional hours fall below 1040. Given current scheduling, this amounts to less than three cancellations for most districts. Conveniently (for the purposes of this study), Colorado administers its standardized tests in March, months before any missed days are ever made up. The same can be said for Maryland, which administers its tests at the end of April while school releases in June.

Because histories of cancellations for Maryland schools are not maintained, Marcotte estimates the reduced form relationship between aggregate snowfall and student performance. Marcotte and Hemelt (2007) obtain partial cancellation histories for Maryland, finding instructional days have significant effects on performance. However, due to the incomplete nature of cancellation histories they pool together two testing regimes (MSPAP (1993-2002) and MSA (2003-2005)) which introduces a potential selection problem: districts with relatively few cancella-

tions tend to not maintain cancellation histories as far back as other schools.[2] Colorado also fails to collect closure histories in any unified location. I overcome this obstacle for both states by using a two sample estimation technique similar to two-sample IV (Angrist and Krueger, 1995). For the 06/07 and school year I have collected daily cancellation information as reported by web sites and news agencies in Colorado and Maryland, thereafter verified through calls to school districts. Using climatic data, one can approximate the structural relationship between snowfall and cancellations. Combining a first stage of weather's impact on cancellations and a reduced form of weather's relationship with student performance, school cancellations' effect on student performance is inferred through indirect least squares. This allows one to study the effect of weather-related cancellations over long periods of time, even if cancellation histories are not maintained. Using this approach, future studies can easily confirm the effect weather-related cancellations on student outcomes, even if limited information on cancellations is available.

A second identification strategy investigates test examination dates, which changed 5 times over 5 years in Minnesota. The changes in test dates alternated between moving the test earlier and later. They were moved earlier by 10 and 11 school days (in 2002 and 2004), and were scheduled later by 10 school days and 15 school days twice (in 2001, 2003, and 2005 respectively). This created

[2]Marcotte and Hemelt adjust for this using district-specific time trends.

substantial variation both increasing and decreasing the amount of time students received prior to examination.

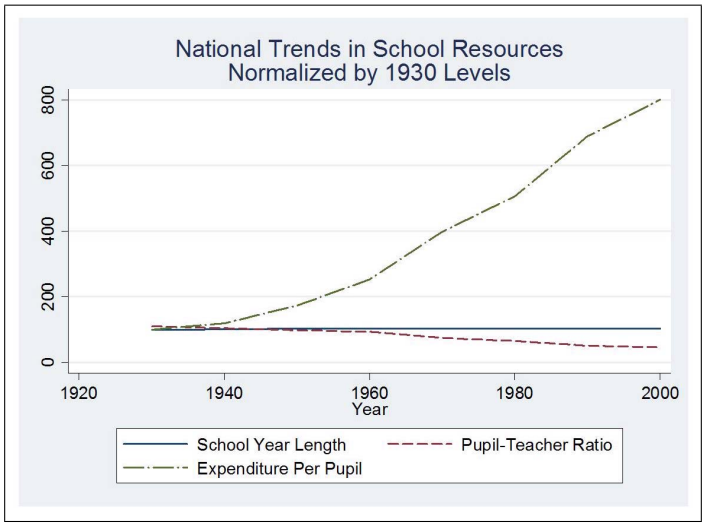
Both sources of variation yield similar evidence regarding school year length's effect on student performance. Because the available performance variables are proportions, the effects are estimated using familiar probability models for grouped data. In addition, because the latent variable is a test score, the estimated effects on the latent variable have a valuable economic interpretation: how many standard deviations average scale scores have changed. Both the response probabilities and the implied effect on latent test scores yield evidence that increased instructional days raise student performance. The evidence suggests that extending the school year can be a method of increasing student performance, and perhaps with it, human capital accumulation.

2.2 School Year Length: Background and Identification

The education production function is a common model used to study the choices of administrators and their ultimate consequences. The administrators are free to pick the levels of various inputs in the educational process, subject to their budgets and state guidelines. Examples of inputs include teachers (in number or quality), textbooks, and the length of instruction time. Outputs of the

educational process include test scores, grades, graduation, going to college, and finding jobs, among many others. Figure 2.1 compares the national trends of the pupil-teacher ratio and real per pupil spending against school year length over the last century. Contrasting the trends, expenditures on teacher employment have risen considerably, while little funds have been devoted to extending the amount of instructional time students receive. If longer school years do improve student outcomes, they could be an alternative to other policies that influence school quality.

Figure 2.1: Trends in Education Inputs, 1930-2000



The magnitude of school year length's impact on student outcomes is largely an empirical question. However, comparing across states or nations to assess school year length's effect can introduce problems of bias. Actual instructional days can be divided into two parts: the planned instructional days and cancel-

lations. Planned instructional days are under the control of the administrator, subject to budgetary constraints and time. Most previous studies have focused on differences in planned instructional days, identified by comparing across states or nations. However the differences in planned instructional days can be largely due to differences in budgets, introducing possible upward bias. Also one might have concern that struggling schools might extend their school year to improve performance on standardized tests, which would bias school year length's effect downward. Texas recently required all districts to begin every year on the last Monday in August for this reason.[3] Using planned instructional days can bias school year length's effect, and the sign of the bias is arguably indeterminate. Thus studying the component of school year length under the control of administrators – planned instructional days – may be counter productive.

Using variation in instructional days due to weather-related cancellations can eliminate the selection problems associated with longer planned school years (which could indicate greater school resources or poor performance on prior exams). The part determined outside the control of administrators still informs about the general effects of increasing instructional days, as weather-related cancellations reduce the amount of time teachers have to instruct, quiz, or meet with students.[4]

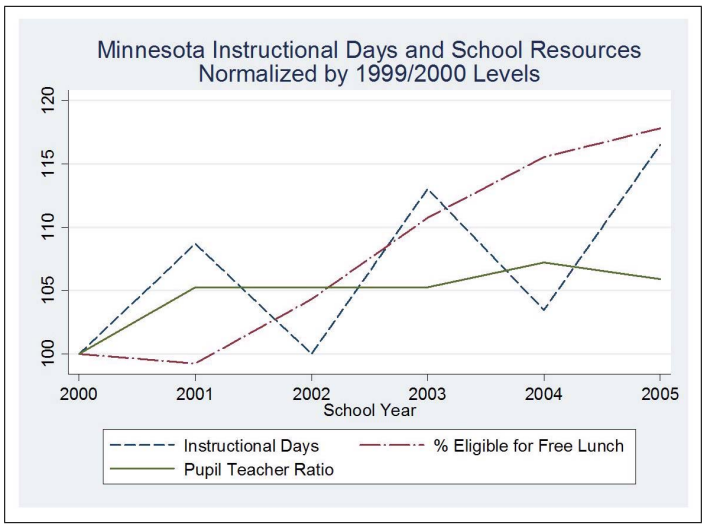
Cancellations due to weather identify the effect of instructional days based on yearly fluctuations due to weather, when the test date is fixed. Another possibility

[3]Dallas Morning News, Thursday May 4, 2006.

[4]We discuss in Section IV reasons weather-related cancellations could under or overstate school year length's causal effect.

is to study situations were dates of examination are shifted. This approach could share some common advantages with weather related cancellations. The variation occurs within districts, and changing the date of test administration does not alter other school resources. Although schools might wish to move their date of examination for endogenous reasons, in Minnesota all the shifts were statewide. In addition, from 2000 to 2005 the date of examination alternated being shifted later and earlier. So trends which are relatively smooth – such as changes in demographics or school quality – can be accounted for. Because a change in instructional time due to a test date shift is known at the beginning of the school year, teachers have time to plan out their year accordingly. For this key reason, examination date changes may more closely resemble an extended or shortened school year.

Figure 2.2: Minnesota Resource Trends



2.2.1 Exogeneity of Weather: Snowfall's Spatial Distribution

A critical assumption in order for cancellations due to weather to identify the causal effect of instructional days is that cancellations be randomly assigned to schools. Even though weather is exogenous, if it is correlated with unobserved elements that impact student performance, causal effects remain unidentified. Thus choosing the correct sample framework, cross-section or panel, can be vital to identifying a causal effect.

Snow accumulates heavily along the mountain range in the middle of Colorado, and neglects to impact the southeastern region. Income in Colorado follows nearly the same spatial pattern. Though not as clear as Colorado's, it seems the correlation between snowfall and income is reversed in Maryland, with the poorest regions in the western strip of Maryland receiving the most snow while the wealthiest regions receive only mild amounts.[5] Although snowfall is exogenous, choice of residence is not random throughout the two states.

To see the extent of correlation between snowfall and resources in Colorado and Maryland, cross-section regressions are estimated using weather as the dependent variable. These are done purely to measure correlation between levels of weather and levels of resources, with the results presented in Table 2.1. In Colorado, districts with substantial snowfall tend to be rich districts while the

[5]Figures in the appendix demonstrate the spatial patterns across the two states.

correlation between snowfall and student family income varies by year in Maryland. However running the regression as a panel and controlling for district fixed effects and year specific trends, none of the variables are by themselves or jointly significant. So although snowfall exhibits spatial correlation with student or school resources, schools experiencing variation in snowfall are not systematically expe-riencing changes in school resources. Controlling for school level fixed effects and yearly trends can eliminate the selection bias that would be introduced due to non-random selection of residence in Maryland and Colorado.

2.2.2 Minnesota: Examination Date Variation

Another source of variation in instructional days exploited is the shifts in scheduled test date administration for the Minnesota Comprehensive Assessment. Minnesota is one of six states which mandates that school start after a specific date, with the remaining states leaving it to the discretion of local school districts.[6] Its September 1 starting date is also tied for the latest.[7] Between the years 2000 and 2005, the Minnesota Department of Education moved the date for its assessment each year, and by several days each time. Because of the shared mandated starting time for schools, shifts in the test date create the potential for more or less instructional time. The trend of average test scores is plotted against the number of instructional days prior to examination in Figure 2.3. Every time

[6]The other five are Texas, Indiana, North Carolina, Virginia, West Virginia. Taken from Education Commission of the States.

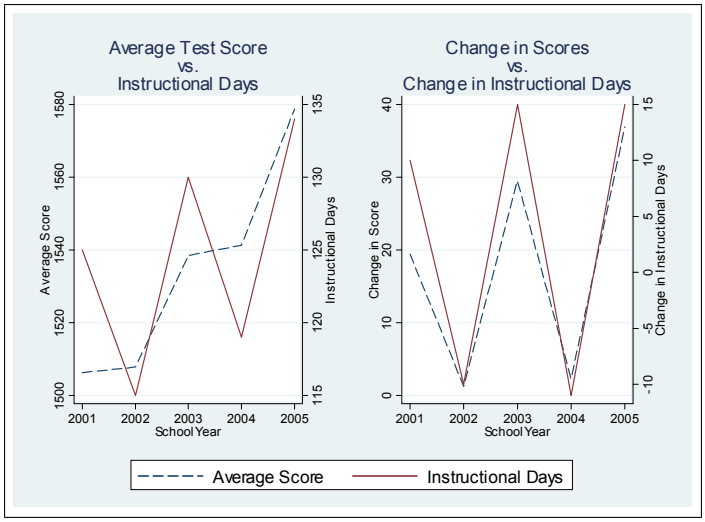
[7]Minn. Stat. 120A.41. Also in consequence, most schools begin the day after labor day.

Table 2.1: Correlation Between Snowfall and Resources

Year	99/00	00/01	01/02	02/03	03/04	Panel
Colorado						
Pupil Teacher Ratio	.	.	.28**	.17	.12	-.008
			(.14)	(.15)	(.08)	.159
% Reduced Price Lunch	.	.	-17.26***	-12.99***	-12.99***	1.79
			(4.62)	(4.11)	(4.11)	(3.034)
Fixed Effects	No	No	No	No	No	Yes
Year Dummies	No	No	No	No	No	Yes
F-Test	.	.	15.32***	7.92***	8.53***	0.17
Maryland						
Pupil Teacher Ratio	-.041	.054	.025	.	.	-.011
	(.043)	(.0493)	(.020)			(.018)
% Reduced Price Lunch	-29.87	-35.99	-12.98	.	.	1.027
	(35.74)	(34.67)	(14.40)			(14.05)
Fixed Effects	No	No	No	No	No	Yes
Year Dummies	No	No	No	No	No	Yes
F-Test	0.53	0.72	1.65	.	.	.17

Dependent Variable: Yearly Snowfall. All results clustered at district-year level.
***significant at 1%, ** significant at 5%, % significant at 10%. Parentheses indicate standard errors.

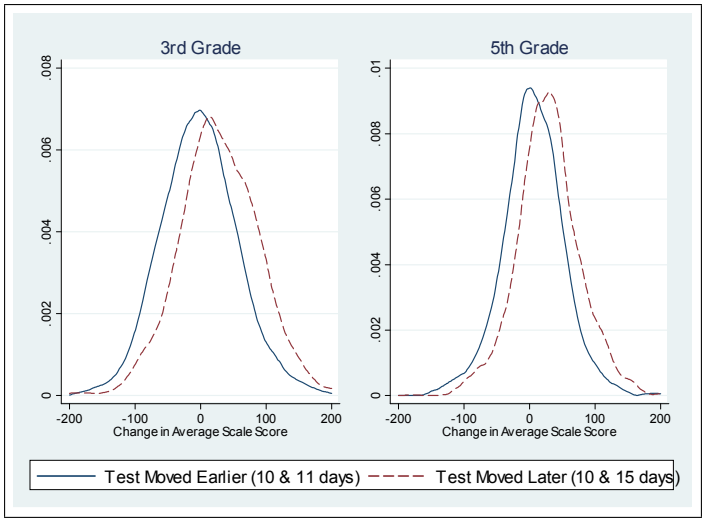
Figure 2.3: Test Score and Instructional Day Trends



the test date is moved earlier, the trend flattens out, while tests administered later in the year show considerably more improvement. This is mirrored when plotting the change in average test scores against the change in instructional days.

The same effects are observed at a more disaggregated level. Using school level average test scores in Figure 2.4, I plot the distribution of the change in average scale scores for years with tests earlier in the school year, contrasted with the distribution of the change in scores for tests administered later in the year. The year-to-year change in scale scores is shifted to the right for both grades when the test is administered later in the year.

Figure 2.4: Distributional Shifts from Early vs. Late Test Dates



2.3 Specification and Estimation

The student performance data are results from the Colorado, Minnesota, and Maryland State Assessments. Each of the tests has stakes for teachers and administrators, but not for students.[8] Mathematics exams are studied because they are relatively objective and cover a consistent curriculum. All 3 states publicly make available grouped averages of performance, which will be the dependent variables of interest when calculating school year length's effect. However, it is useful to consider a simple model of student performance at a micro level to accurately interpret the results and establish identification.

[8]Depending on how close students are to a threshold they may exhert more or less effort to pass an exam, which has stakes for the student. See Betts (1996).

2.3.1 Micro Model of Student Performance

Consider a model of testing where a student's performance depends on his or her observable characteristics, his or her school's resources, and instructional days. Instructional days are the planned instructional days less cancellations, where cancellations are influenced by weather (snowfall in particular) and planned instructional days depend on resources. With information on individual student test scores, one could estimate linear regressions for the following model.

$$T_{ist} = I_{st}\beta + X'_{it}\beta_X + R'_{st}\beta_R + s_s + \tau_t + \varepsilon_{ist}$$

$$I_{st} = P_{st} - C_{st}$$

$$P_{st} = R'_{st}\alpha_R + e_{st}$$

$$C_{st} = w_{st}\alpha + v_{st}$$

T_{ist}	:	Student $i's$ test performance at school s at year t
I_{st}	:	Actual Instructional days for school s at year t
P_{st}	:	Planned instructional days for school s at year t
C_{st}	:	Cancellations for school s at year t
X_{it}	:	Characteristics for student i at year t
R_{st}	:	Resources at school s at year t
w_{st}	:	Weather for school s at year t
s_s	:	School fixed effect
τ_t	:	Year fixed effect

At this point a student's performance depends on resources, both at the individual and school level, and instructional time. The reduced form impact of weather on student performance is

$$\frac{dT_{ist}}{dw_{st}} = -\beta\alpha.$$

Ideally, we would construct a weather measure w_{st} such that $\alpha = 1$.[9] If $\alpha = 1$, then there is no need for a first stage as using the reduced form estimates of weather would be equivalent to the structural relationship instructional days and student performance.[10] Aggregate snowfall is likely to be correlated with closures but can be improved upon.[11] For instance 10 days where it snows 1 inch will probably not lead to any cancellations. However, one day with 10 inches of snow almost surely would. Aggregate snowfall would treat these realizations of weather the same. To more accurately assess weather likely to cancel school, I also construct measures of weather based on the number of days on which snowfall exceeded thresholds. Of course another trade-off exists. Thirty inches of snow on one day might cancel school for the next 3 or 4 days, but threshold variables would treat this as equivalent to one day with 4 inches of snow. For completeness, both weather measures are considered.[12] Consider now the reduced form representation, removing instructional days directly from the regression.

$$T_{ist} = w_{st}(-\beta\alpha) + X'_{ist}\beta_X + R'_{st}(\beta_R + \beta\alpha_R) + \beta v_{st} + s_s + \tau_t + \varepsilon_{ist}$$

One can rewrite the expression above, getting

[9] In a slight abuse of notation I refer to different weather measures having different α's. This could be more accurately respresented as $\alpha_{w_{st}}$. To avoid more cumbersome notation, we will refer to them all as α.

[10] The reduced form for any weather measure could be rescaled so that $\alpha = 1$. Essentially a first stage relationship tells one how to rescale the units on the weather variable so that $\alpha = 1$.

[11] Marcottes chosen regressor. Marcotte uses yearly snowfall, I remove snow during winter vacation, weekends, or school holidays.

[12] There are many other weather measures that could be used. However, any measure highly correlated with cancellations that doesn't impact students other than through cancellations is sufficient and necessary for identification.

$$T_{ist} = w_{st}\gamma + X'_{ist}\beta_X + R'_{st}\psi + s_s + \tau_t + u_{ist}, \tag{2.1}$$

where $(\beta e_{st} - \beta v_{st} + \varepsilon_{ist}) = u_{ist}$, $-\beta\alpha = \gamma$, and $(\beta_R + \beta\alpha_R) = \psi$.

This regression would be easy enough to run, were micro-level data on student performance available. However, state assessment results made publicly available contain grouped information for grade levels within schools. Maryland publishes the proportion proficient and advanced while Colorado releases the proportion partially proficient, proficient, and advanced. Minnesota reports the proportion partially proficient, proficient, advanced, and also average test scores.[13] One can still estimate the effects on student performance, but the data requires it be done in the context of probability models.

Then if we are interested in the probability that $T_{ist} \geq t^*$, where t^* is an academic standard, the partial effects include the reduced form effect on test scores, along with the density at the cut off point. To illustrate this point, let us rewrite the effect of weather on student performance, given the data refer to the probability of exceeding an academic standard.[14]

[13]Maryland calls its middle category satisfactory, while Colorado and Minnesota refer to it as proficient. Maryland calls its highest category excellent, while Colorado and Minnesota name it advanced. In this paper, satisfory and excellent proportions in Maryland will be refered to as proficient and advanced for simplicity.

[14]We continue with the representation of a micro probability model. Grouped probability estimates have similar similar response probabilities, controlling for average student characteristics rather than particular traits.

$$P(T_{ist} \geq t^*)$$

$$= P(w_{st}\gamma + X'_{ist}\beta_X + R'_{st}\psi + s_s + \tau_t + u_{ist} \geq t^*)$$

$$= P(u_{ist} \geq t^* - (w_{st}\gamma + X'_{ist}\beta_X + R'_{st}\psi + s_s + \tau_t))$$

$$\Longrightarrow$$

$$\frac{dF}{dw_{st}} =$$

$$\underset{\text{(I)}}{f(t^* - (w_{st}\gamma + X'_{ist}\beta_X + R'_{st}\psi + s_s + \tau_t))} \quad \underset{\text{(II)}}{(\gamma)}$$

or

$$\frac{dF}{dw_{st}} =$$

$$\underset{\text{(I)}}{f(t^* - (w_{st}\gamma + X'_{ist}\beta_X + R'_{st}\psi + s_s + \tau_t))} \quad \underset{\text{(II)}}{(-\beta\alpha)} \quad .$$

The partial effects have two components: the effect of weather on latent test scores (II), and the density at the cutoff (I). One can estimate the impact at different academic standards, grades, or demographic groups. However, differences in the response probabilities reflect both variation in the latent effect on student performance or different densities of students at the chosen standard. In addition, typical procedures used to scale effects can be problematic in probability models.

When interpreting the practical size of coefficients, one method is to compare the estimated effect to the mean of the dependent variable. This can give the researcher an approximation of the "percentage" effect. With probability models this is problematic because the choice of failure and success is arbitrary. For this problem, I could have chosen to examine the probability of being below an academic standard. The new estimated effect is the same in magnitude, only with

the sign reversed. However, the proportion of students below the standard is by definition one less the proportion above. By comparing the partial effect to the mean of the dependent variable, one can either inflate or deflate the "percentage" effect depending on the arbitrary definition of success and failure.[15] In other words, using the mean to scale the effect is not invariant to the researcher's choice of success in probability models. This obstacle can be partially overcome depending upon which probability model is implemented.

I proceed now to the estimators used in grouped probability models. Probability is replaced by its sample next-of-kin, the proportion of the students exceeding a threshold. Minimum chi-square methods provide several different well studied estimators from which to choose.[16] I examine the linear probability model and normit presented below (with a general dependent variable and vector of regressors).

Linear Probability Model	$P_t = X_t'\beta + u_t$
Normit	$\Phi^{-1}(P_t) = X_t'\beta + u_t$

The linear probability model is familiar from binary outcomes, and the normit, the grouped version of a probit, is a reasonable choice for test scores well-approximated by a normal distribution.[17] Also the β estimated by the normit regression has a valuable interpretation. The original dependent variable is bounded

[15] This is similar to estimating elasticities. We could use the initial or the end point to scale the change. For this reason it is common to used the midpoint to get an average elasticity. In our case, the midpoint is always .5 by definition of probability.

[16] See Madalla for an extensive chapter on micro and grouped probability models (1983).

[17] Typically there is some skewing in tests. Early grades they are skewed right, and scores skewed left for later grades. Grades in the middle are typically those most symetrically distributed.

between zero and one. When taking the normit (inverse cumulative normal) transformation of the proportion variable, the transformed variable is a standard normal variable. Due to the normit transformation, the estimated β indicates how many standard deviations latent test scores have shifted. Focusing on the impact on the latent variable (which is called the the latent effect from this point on in the paper) rather than response probabilities also provides an invariant way to compare partial effects, as the transformation eliminates the density component. This is one of the few situations where the untransformed coefficient in a probability model has a valuable economic interpretation. This is useful for comparing effects across grades, proficiency standards, or states, which have both different standards and densities of students.[18]

2.3.2 First Stage: Weather and Cancellations

Up to this point, the focus of the discussion has been on estimating the reduced-form effect of weather on student performance, γ. In order to place a magnitude on how additional school days affect student performance, the reduced-form effect needs to be scaled by the relationship between weather and cancellations, α. Because cancellation histories are not maintained, this data challenge is overcome by estimating a first stage equation for the 2006/2007 school year. The weather

[18]If the data are truely generated by a normal distribution and the effect of instructional days is linear, then the latent effect will be the same across across thresholds. If the effect differs across performance standards, this could be both due to non-normality or non-linearity of the effect.

variables previously discussed are included as regressors in the first stage regressions. Two possible specifications for a first stage are explored. A high frequency approach estimates how well the weather variables predict closures on a particular day. A low-frequency analysis estimates how weather over the course of the school year predicts the number of cancellations occurring within that year.

The first specification's dependent variable is an indicator for whether school is open or cancelled at a particular district on a given day. Because the data are measured at the daily level, there are likely to be some matching problems. For example, snowfall on a Monday night would cancel school on Tuesday but is matched with Monday's school closure status. This measurement error will likely attenuate α towards zero. In addition, because the threshold variables are indicators, α will be naturally bounded between zero and one. Because β is the parameter of interest, attenuation of α would bias the estimate of β away from from zero as the reduced form effect γ is divided by -α to recover β (cancellations refer to lost days, so dividing by -α yields a β that corresponds to the effect of an additional day of schooling). For this reason, it may useful to think of the indirect least squares estimates as upper bounds.

$$Cancellation_{sd} = \alpha_o + \alpha w_{sd} + d_s + v_{sd}, \tag{2.2}$$

where s indicates district and subscript d denotes the day and d_s is a district fixed-effect.

The low frequency approach uses the number of cancellations as the dependent variable, aggregating equation (2.2). The true population parameters remain unchanged with this aggregation for the population model, due to linearity. However misclassifications of weather due to calendar effects may be reduced as a lot of snow on Monday evening or Tuesday morning would both aggregate to 1 day with a lot of snow.[19]

For completeness, both the low-frequency and high-frequency methods to estimate α are computed and if there is little mismatching of the weather variables, the estimates will be similar. Regardless, after estimating α the standard errors need to be adjusted to account for both the randomness of $\hat{\gamma}$ and $\hat{\alpha}$. The limiting distribution of $\frac{\hat{\gamma}}{\hat{\alpha}}$ is approximated using the delta-method. Recall for an estimated parameter vector $\hat{\theta}$, $g(\hat{\theta})$ has the following limiting distribution where $G(\theta)$ is the matrix of partial derivatives with respect to θ.

$$N(g(\hat{\theta}), G(\theta)'V(\hat{\theta})G(\theta))$$

In our case, the form of $g(\hat{\theta})$ is $\frac{\hat{\gamma}}{\hat{\alpha}}$.[20] The reduced forms for Colorado and Maryland are estimated respectively for 2002-2006 and 1993-2002. For both states

[19]Notice for either specification, the other controls have been ommitted from the first stage. This is mainly due to the fact that the regressors are not yet available for the 2006/2007 school year. In addition, in the high frequency approach any variables that are time constant are absorbed because of the fixed effects. Because this includes any regressors that don't vary throughout a school year, the fixed effects are collinear with all school and student characteristics recorded at the yearly level.

[20]This typically requires continuity of the function of the parameters. This function is continuous everywhere, except where $\alpha = 0$. This is a common problem of exactly identified instrumental variables equations. In the results section the first stage is sufficiently powerful to reject the null that $\alpha = 0$.

the first stage is estimated for the 2006/2007 school year. Because the parameters are estimated from separate samples, it is assumed that the off-diagonal elements of the variance-covariance matrix are zero.

$$g(\gamma, \alpha) = \frac{\gamma}{\alpha}$$
$$G(\gamma, \alpha) = \begin{pmatrix} \frac{1}{\alpha} \\ \frac{-\gamma}{\alpha^2} \end{pmatrix}$$

$$\frac{\hat{\gamma}}{\hat{\alpha}} \rightarrow^d N\left(\frac{\gamma}{\alpha}, \frac{var(\hat{\gamma})}{\alpha^2} + \frac{var(\hat{a})\gamma^2}{\alpha^4}\right) \tag{2.3}$$

Notice if $\alpha = 0$, the mean and variance will be infinite, making the distribution undefined. This makes a powerful first stage critical to this study, like any instrumental variables approach.

2.3.3 Minnesota

For Minnesota, similar regressions are estimated, albeit without some of the complications using weather to generate random variation in instructional days. The regressor of interest is simply the number of days prior to examination. This is found by calculating the number of potential school days between the first of day of school and the test date (removing holidays, weekends etc.). Because historical school schedules are not maintained, winter break is defined to be between December 23 and January 3. Though there might differences in winter break

length, the fixed effects will capture any time constant discrepancies. So even if some schools have more instruction (due to winter break differences) than others, the deviation in instructional days from the mean will be the same for all school districts. If there are changes in winter break length over time (or weather-related cancellations), this would introduce measurement error, attenuating the estimates. One caveat is that because schools are experiencing the same deviation from their mean instructional time, instructional days would be correlated with year effects. In order to adjust for trends (which Figure 2.3 strongly suggests exist), school specific quadratic trends are included in the regressions.

2.4 Results

2.4.1 Data Sources

The performance data are taken from mathematics results made publicly available from the Maryland, Minnesota, and Colorado Departments of Education. The Maryland assessment results are from 1993-2002, Colorado's cover 2002-2006, and Minnesota's span 2000-2005. The 3rd, 5th, and 8th grades are studied in Maryland, the 8th grade is explored in Colorado, and the 3rd and 5th grades are examined in Minnesota.[21] Maryland and Minnesota also make available the

[21] In Colorado, schools following a year-round schedule or 4-day school were excluded. This because details regarding breaks for year-round schools were not maintained, and 4-day schools report which weekday they have off since 2003, but not prior. Although the exam began administration in 2000, 2002 on is studied because the scale scoring changed to a new regime in 2002.

variables used as controls, while the control characteristics for Colorado are taken from the National Center of Educational Statistics. The weather data are daily surface observations from the National Climatic Data Center (details contained in the appendix on the linkage). Data on cancellations for Colorado and Maryland were obtained by surveying school districts at the end of the 06/07, details of which can be found in the Appendix.[22] The summary statistics are found in Table 2.2.

Control characteristics of the schools and their student bodies are included in the regressions. Even though weather and the test date changes are plausibly exogenous events, including the controls can prevent spurious correlation and also reduces sampling error. Common controls to all regressions run include the fraction of students eligible for reduced price lunches and the pupil teacher ratio. School fixed effects are included in all regressions, while year dummies account for trends in Colorado and Maryland and quadratic school specific trends are included for Minnesota. Maryland and Minnesota have a few unique controls not available through the National Center of Educational Statistics.[23] Colorado and Maryland also both report information on teaching assistants per pupil. Maryland and Minnesota both record the proportion of students which are limited-English proficient. Maryland provides yearly data on per capita wealth and the fraction

[22]In Colorado 107/178 districts provided cancellation for 2006/2007 school year, while in Maryland 19/24 responded.

[23]For Colorado, controls for the 05/06 school year had not been released yet. The prior years values were imputed for these missing obsevations. Also the mean of previous values was tried. The results are robust to method of imputation, or excluding the controls.

Table 2.2: Summary Statistics

Variable	Maryland	Colorado	Minnesota
% Advanced	.10	.14	.15
	(.14)	(.11)	(.11)
% Proficient	.44	.41	.55
	(.22)	(.18)	(.17)
% Partially Proficient	.	.72	.71
		(.16)	(.15)
Snowfall	11.55	20.26	.
	(13.55)	(10.80)	
Pupil Teacher Ratio	16.5	16.3	16.0
	(1.63)	(2.31)	(1.81)
% Eligible for Free Lunch	.33	.37	.33
	(.25)	(.24)	(.22)
Teaching Assistants/1000 Students	10.91	17.97	.
	(3.63)	(7.82)	
Average Teacher Experience	.	.	16.85
			(4.91)
Median Wealth Per Student	223,785	.	.
	(84,966)		
% Title One	.15	.	.
	(.32)		
% Limited English Proficient	.020	.	.054
	(.04)		(.11)

Parentheses indicate standard errors.

of students which are Title I eligible and Minnesota has information on the average experience of teachers. Excluding or including these additional variables in Maryland or Minnesota has little impact on the results. Also, all the reduced-form regressions are weighted by the number of students taking the test, and because the level of snowfall is shared by all schools within a district in a year, standard errors are clustered by district and year.[24]

2.4.2 First Stage Estimates

To infer the effect of additional instructional days on student performance, weather's reduced form effects need to be scaled by weather's relationship with cancellations for Colorado and Maryland. The high frequency approach employees a linear probability model and includes district level fixed effects. An observation is a day for a district. Meanwhile, the low frequency approach aggregates over the year and compares across districts. The results are below in Table 2.3.

Both approaches yield similar estimates for Maryland. Each additional inch of snowfall increases the odds of a cancellation by .16. The high frequency estimate is precise enough with an F-statistic of 20.15 to suggest the instrument is not weak. This suggests that the reduced form coefficients should be scaled up by a factor of 6 in Maryland. Colorado superintendents are more resistant to snow, as an additional inch of snow is estimated to raise the probability of cancellation by

[24]This is in part due to relatively few districts in Maryland. Only 24 exist, and some are quite large with dozens of schools. Thus if one clustered at the district level, some clusters could take up excessively large partitions of the variance-covariance matrix.

Table 2.3: First Stage, Effect of Snowfall on Cancellations

	High Frequency		Low Frequency	
Weather Measure	Colorado	Maryland	Colorado	Maryland
Snowfall	.052***	.16***	.013	.14**
	(.004)	(.03)	(.010)	(.056)
t-statistic	11.52	4.49	1.33	2.5
f-statistic	132.80	20.15	1.76	6.23
# Days Snow > 4 inches	.37***	·	.27**	·
	(043)		(.109)	
t-statistic	8.66	·	2.68	·
f-statistic	75.01	·	7.18	·
# Days Snow >1 s.d.	.37***	·	.23**	·
	(.044)		(.11)	
t-statistic	8.44	·	2.05	·
f-statistic	71.27	·	4.21	·
Fixed Effects	yes	yes	no	no
Number of Observations	17716	2681	107	18

***significant at 1%, ** significant at 5%, * significant at 10%.

Results clustered at district-year level. Parentheses indicate standard errors.

.05, somewhat smaller than in Maryland. For every day with snow greater than 4 inches, the probability that Colorado school districts cancel school increases by .37. The high frequency regressions provide the most precise estimates of the structural relationship between weather and cancellations, all passing weak instrument standards, and hence are used for the indirect least squares estimates of an instructional day's effect. Any measure of weather could be linked with its reduced form for Colorado. The number of days with snow greater 4 inches is used for Colorado and inches of snowfall is used for Maryland for the final estimates presented in the next section, and the results are similar across other weather measures.[25]

[25]The 4 inch threshold measure seems the most robust across the two frequencies. In addition,

2.4.3 Reduced Form Estimates

Both the linear probability model and normit will be used in estimating the reduced form effect of snowfall and performance. Recall $\frac{dF}{dw_{st}} = f(\cdot)(-\alpha\beta)$.If one uses the same weather variable and compares across performance measures, differential effects could reflect both differences in effects on latent performance β, as well the density of students at the cutoff. Because the density is always greater than or equal to zero we can identify the sign of $-\alpha\beta$ above, but relative magnitudes cannot be compared because of differences in densities. Two model specifications are used to estimate the response probabilities, and the effect on latent scale scores. The linear probability model is used in estimating response probabilities, as it does not require specification of $f(\cdot)$ thus offering some additional robustness properties.[26] The untransformed normit coefficients will provide estimates of the effect of weather on latent scale scores.

As shown in Table 2.4, the proportion above each of the academic standards falls more days with substantial snowfall. In addition, the effects are strongest low in the test score distribution, as the impacts of all the weather variables on the proportion partially proficient are larger and more statistically significant than the effects at other proficiency cutoffs.[27] With each day of snowfall with more

it may be less sensative than snowfall to outliers such as the large snow-storm which hit Colorado December 20-23, 2006.

[26]It should be noted that estimated normit response probabilities closely mirror those estimated by the linear probability model and are available upon request.

[27]This could also be due to a local effect, if districts that experience the most variation in snowfall are also those whose density is most concentrated around the partially proficient standard. See Angrist and Imbens (1994).

Table 2.4: Colorado Reduced Form

Effect of an Additional Day with Snow>4 inches on Performance

Dependent Variable	Grade 8
Response Probabilities	
% Partially Proficient	$-.0056^{**}$
	(.0024)
% Proficient	$-.0030$
	(.0032)
% Advanced	$-.0043^{**}$
	(.0021)
Shift of Scale Scores in Standard Deviations	
Latent Partially Proficient	$-.015^{*}$
	(.008)
Latent Proficient	$-.0053$
	(.009)
Latent Advanced	$-.0073$
	(.009)
Fixed Effects	Yes
Controls	Yes

***significant at 1%, ** significant at 5%, * significant at 10%.

All results clustered at district-year level. Parentheses indicate standard errors.

than four inches, the fraction partially proficient declines by .0056 From the normit regression, an additional day with snow greater than 4 inches decreases test scores by .015 standard deviations (at the partially proficient standard). The direction of the effects is clear, increases in snow is associated with lower student performance.

Maryland shows similar results to Colorado's, albeit with greater statistical precision. The results are presented in Table 2.5. Rows labeled "proportion proficient" and "proportion advanced" refer to the linear probability estimates, while "latent proficient" and "latent advanced" refer to the untransformed normit coeffi-

Table 2.5: Maryland Reduced Form

Effect of An Additional Inch of Snowfall on Performance

Dependent Variable	Grade 3	Grade 5	Grade 8
Response Probability			
Proportion Advanced	−.000052	−.00048***	−.00046***
	(.00013)	(.00015)	(.00016)
Proportion Proficient	−.00050	−.00073**	−.00053**
	(.00038)	(.00032)	(.00022)
Shift of Scale Scores in Standard Deviations			
Latent Advanced	−.00049	−.0027***	−.0021***
	(.00098)	(.00081)	(.00068)
Latent Proficient	−.0018	−.0024**	−.0014**
	(.0012)	(.0010)	(.0006)
School Fixed Effects	Yes	Yes	Yes
Controls	Yes	Yes	Yes

***significant at 1%, ** significant at 5%, * significant at 10%.

All results clustered at district-year level. Parentheses indicate standard errors

cients.[28] An additional inch of snowfall decreases the proportion scoring proficient by .00073 for the 5th grade and .00053 in the 8th grade. Likewise, an inch of snowfall is estimated to decrease latent scale scores by .0024 standard deviations (for the fifth grade at the proficient standard). With the exception of the third grade, the estimated effects are all significant. Once again, increased winter weather, in the form of inches of snowfall, is associated with reduced performance for all grades and proficiency levels.

[28]Because the normit function is not defined for proportions equal to zero or 1, these are replaced with small deviations, i.e. 0.01 and .99.

2.4.4 Final Estimates of the Effect of Additional Instructional Days

Because Colorado and Maryland's indirect least squares estimates refer to the effect of losing an instructional day, those estimates are multiplied by -1.[29] With this slight transformation in mind, Table 2.6 compares the estimates of the effect of an additional day of schooling for all three states across various grades and thresholds of proficiency. Rows with proportion variables are estimated using linear probability models, while rows denoted as latent refer to untransformed normit coefficients. All have similar qualitative implications: additional instructional days improve student performance. Most are highly significant, though there are some differences in magnitude.

Because the density of students varies across grades and academic standards, the best measures to compare across grades and states are probably the latent effects. The estimated effects derived from weather-related cancellations are in general larger than those from test-date changes. In Maryland, an additional day of schooling is estimated to improve test scores by as much as 0.016 standard deviations, while an additional day improves test scores by at as much as 0.013 standard deviations in Minnesota. For Colorado, the largest estimate suggests an

[29]This is because the fixed effect regressions refer to deviations from means, the estimated coefficents can refer to deviation above the mean (more snow days) or below (less snow days). Linearity of the regression model allows this transformation.

Table 2.6: Final Estimates of the Effect of an Additional Day of Schooling

	Colorado	Maryland			Minnesota	
Dependent Variable	Grade 8	Grade 3	Grade 5	Grade 8	Grade 3	Grade 5
Response Probability						
Proportion Advanced	.011**	.00038	.0030***	.0029**	.0028***	.0011***
	(.0055)	(.00081)	(.0011)	(.0012)	(.00027)	(.00016)
Proportion Proficient	.0081	.00031	.0045**	.0033**	.0045***	.0031***
	(.0084)	(.0025)	(.0022)	(.0015)	(.00025)	(.00020)
Proportion Partially Proficient	.014**	.	.	.	.0022***	.0020***
	(.0069)				(.00023)	(.00017)
Shift of Scale Scores in Standard Deviations						
Latent Advanced	.019	.003	.016***	.013**	.013***	.0042***
	(.024)	(.0063)	(.0062)	(.0051)	(.0012)	(.00067)
Latent Proficient	.013	.011	.015**	.0090**	.012***	.0089***
	(.024)	(.0077)	(.0069)	(.0042)	(.00069)	(.00055)
Latent Partially Proficient	.039*	.	.	.	.0074***	.0070***
	(.021)				(.00071)	(.00061)
Percentage Shift in Scale Scores						
Log Average Score	.	.	.	.	.00082***	.00062***
					(.000085)	(.000064)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes			Yes	
Trend	Year Dummies	Year Dummies			Quadratic By School	

***significant at 1%, ** significant at 5%, * significant at 10%.

All regression clustered at district-year level. Parentheses indicate standard errors.

additional instructional day raises test scores by 0.039 standard deviations.

Several factors could explain these differences. First, the estimates reported for Colorado and Maryland use only cancellations in the first stage regression. If delayed starts and early releases (other events that disrupt class and reduce instructional time) are treated as cancellations in the first stage regressions, the final indirect least squares estimates decrease by 25 percent. Non-linearity could also play a role because of decreasing returns to instructional time (Minnesota had more variation than Maryland, which in turn had more than Colorado). Also the effect of additional school days could vary due to test difficulty, student ability, or teacher quality. A few large snow storms impacted Colorado in 2006/2007, which could have attenuated the first stage, and thereby biased the indirect least squares estimates away from zero. Furthermore, weather-related cancellations may reduce critical review time, whereas a moved test date allows teachers to reschedule their time to allow for proper review. These reasons suggest the estimates of instructional days' effect derived from weather-related cancellations could be considered as upper bounds.[30]

In Minnesota, bias could go in the other direction. Several of the test date changes postponed the test until after spring break. If students forget material while on vacation, the Minnesota estimates could understate the effect of addi-

[30]Teacher absences could be an additional concern (Miller et.al. 2007). Because teacher absences are excluded from my data (due to availability), the indirect least squares estimates would be upward biased. This supports the notion that the estimates due to weather related cancellation can be viewed as upper bounds.

tional day of instruction. Also the later dates may have allowed less time for post-assessment material. This could lead to spill overs reducing the amount of material learned before the fourth grade, which could potentially also affect test scores in the fifth grade. These factors suggest that the estimates due to changes in test-date administration could be thought of as a lower bound for instructional days' effect.[31]

Lastly, both identification strategies refer to the effect of a contemporaneous change in instructional days. In essence, they measure the temporary effects of increasing instructional days for a particular school year. If the school year were permanently longer, there could be positive spill-over effects. For this reason, the effect of a permanent increase in school year length could be greater than those estimated in this paper.

2.4.5 Robustness Checks

I proceed to investigate two robustness checks. As pointed out in the previous section, if the ability to remove snow is improving over time, the indirect least squares estimates would overstate the effect of additional instructional days. Another factor that could play a role is school attendance. If school is not cancelled when a snowstorm hits, students might miss school and fall behind their class-

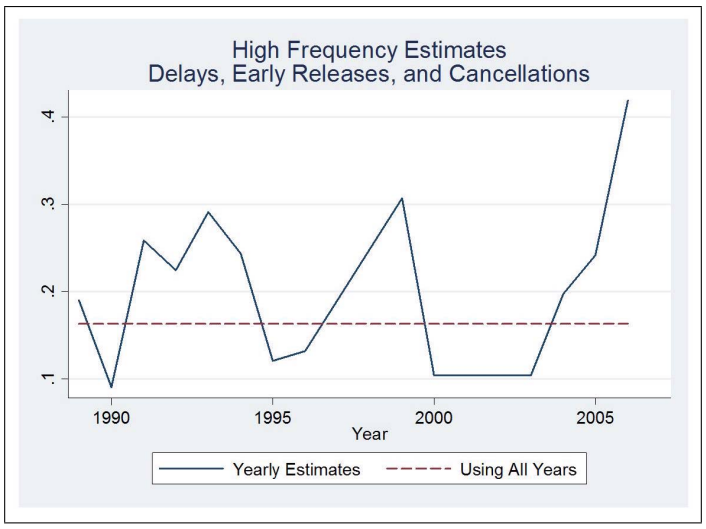
[31] One additional factor that could play a role is absolute age, as students are either older or younger depending on the date of test administration. However students are only older when they take the test, not when they are learning the material throughout the year. For evidence regarding absolute age, see Bedard and Dhuey (2007).

mates. This creates bias as the original reduced form estimates of weather's effect on cancellations would also include the effect of weather on attendance, if there is one. These two sources of possible bias are investigated using additional data sources.

Parameter stability is an implicit assumption of the two sample indirect least squares estimates. If technology in snowfall removal has improved more snow will be required to cancel school, this would bias the indirect least squares estimates away from zero. This concern is likely to be most relevant for Maryland, as the reduced form data go back to the 92/93 school year, while the first stage is estimated for the 06/07 school year. Harford County School District in Maryland has maintained a rich history of weather-related cancellations. From September 1988 through today, they have recorded daily cancellation, delay, and early release information. In addition, total yearly cancellations have been recorded since 1975. These additional data sources offer two ways to test the structural stability of weather's relationship with cancellations. A high frequency analysis estimates the relationship between snowfall and daily cancellations for each school-year beginning in 1988. A low frequency approach will the effect of yearly snowfall on total cancellations for ten year windows, beginning with the 1974-1983 window and ending with 1998-2007 window. Figures 2.5 and 2.6 contain the estimated coefficients for both approaches.[32] Though the estimated relationship between snowfall

[32]The yearly regression could not be estimated in 1997, 1998, and 2002 as there were no cancellations, hence no variation in the depedent variable.

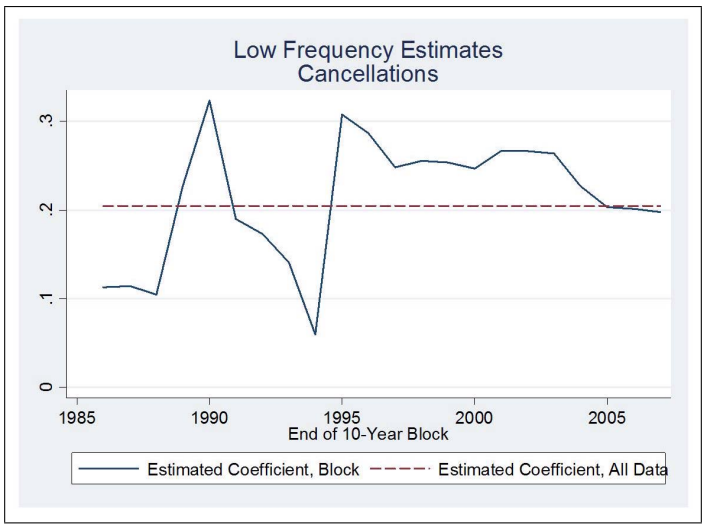
Figure 2.5: High Frequency Stability



and cancellations varies across years, it seems to be noise rather than a systematic trend. Also interestingly enough, the high frequency results from pooling across all years suggest that an additional inch of snow increases the probability of a cancellation by 0.11. This is similar to the earlier results for Maryland using all school districts with only the 06/07 school year.

A last robustness check explores whether snowfall impacts attendance. The hypothesis studied in this paper has concerned the number of days teachers have for instructing their students, not the number of instructional days students choose to attend. If snowfall causes truancy, the indirect least squares estimates will be biased. This is not because the first stage is invalid – rather it concerns the original structural model. Upward bias would occur as part of the reduced form effect is due to attendance, but the current indirect least squares estimates would

Figure 2.6: Low Frequency Stability



attribute it all to cancellations. Beginning in 2005, the Colorado Department of Education has recorded and published the total percentage of hours missed, the percentage of hours missed and excused, and the percentage of hours missed and unexcused. The three measures are regressed on yearly snowfall and the number of days with snow greater than 4 inches in separate regressions–with the results contained in Table 2.7. There is not a statistical or practical relationship between snowfall and the total percentage of hours missed or the percentage of missed hours unexcused.

There is some evidence that greater snowfall increases the number of excused absences. Although the effect is marginally significant, it is small in magnitude. For each day with snow greater than 4 inches, the proportion of hours excused

Table 2.7: Attendance and Snowfall, Colorado

	% of Hours Missed		
	Total	Unexcused	Excused
Snowfall	−.00022	−.0001	.00007
	(.00014)	(.0001)	(.00007)
Days w/Snow>4	.0006	−.0003	.00085*
	(.001)	(.0006)	(.0005)
Fixed-Effects	Yes	Yes	Yes
Mean	.057	.0135	.044

***significant at 1%, ** significant at 5%, * significant at 10%

All results clustered at district-year level

increases by .0008. No matter how it is scaled, the effect is relatively small. In addition it is unknown when the days were missed. Because cancellations require make-up days in the summer, parents could be excusing their students from the make-up days at the end of the school year (because of previously arranged family vacations or other activities). So although there is some evidence of possible bias because winter storms cause excused absences, the correlation between snowfall and excused absences is small in magnitude and could be explained by scenarios that would not bias the results.

2.5 Conclusions

Prior research has been at odds over the effect of school inputs on student outcomes – both labor market and academic. I find evidence that increased instructional days improving student performance. This supports Card and Krueger's findings that longer school years are associated with increased wages. Two differ-

ent identification strategies are employed in calculating the effect of an additional day of schooling, taking advantage of exogenous variation in instruction due to both weather and state mandated shifts in test administration. Also, it is encouraging that the estimates are similar to those of Marcotte and Hemelt. This holds although the method used in Maryland is different along with additional data from Colorado. An entirely different source of instructional day variation in Minnesota provide similar and even stronger results. Weather-related cancellations and test date shifts both offer statistically significant evidence that additional school days increase student performance.

The larger estimates suggest that 5 additional days of instruction would increase test scores by .15 standard deviations, while the smaller suggest it could improve test scores by .05 standard deviations. It may be of use to compare these estimates to other policy interventions, such as decreasing the pupil-teacher ratio. Krueger finds that being in a small class increases a student's percentile ranking by roughly 0.20 standard deviations. This is only a back-of-the-envelope comparison, but it seems that a couple weeks of additional school days is a reasonable substitute for smaller classes.

Although I find evidence of the potential benefits of extending the school year, this does not necessarily justify requiring all schools to do so. In part this is because the costs of lengthening the school year are not homogeneous across districts (due to air conditioning, teacher salaries, transportation). Thus, locations

where it is expensive to lengthen their school year might optimally take advantage other policy interventions, such as reducing the pupil-teacher ratio. This would be consistent with efficient distribution of schooling resources.

In conclusion, my final estimates are consistent with more instructional days raising student performance. Because total instructional days in a school year (pre and post test administration) are fixed despite changes in weather-related cancellations or test-date administration, the estimates relate to the effect of an increase in instructional days. Permanently longer school years could have positive spill-over effects not accounted for by either estimation strategy. The results in this paper suggests longer school years can improve student performance, and perhaps increase human capital accumulated.

Bibliography

[1] Angrist, Joshua and Guido Imbens (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, **62,** 467-476.

[2] Angrist, Joshua D. and Alan B. Krueger (1995). "Split Sample IV Estimates of the Return to Schooling." *Journal of Business Economics and Statistics*, **13**, 225-235.

[3] Bedard, Kelly and Elizabeth Dhuey (2007). "Is September Better than January? The Effect of Minimum School Entry Age Laws on Adult Earnings." *UCSB Working Paper.*

[4] Betts, Julian R. (1998)."The Impact of Educational Standards on the Level and Distribution of Earnings." *American Economic Review*, **88**, 266-275.

[5] Coleman, James S., E.Q. Campbell, C.J. Hobson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Govt. Printing Office.

[6] Eren, Ozkan and Daniel Mittlemet. (2007). "Time to Learn? The Organizational Structure of Schools and Student Achievement." *Empirical Economics*, **32**, 301-332.

[7] Hanushek, Eric A. (1981)."Throwing Money at Schools." *Journal of Policy Analysis and Management*, **1**, 19-41.

[8] Krashinsky, Harry (2006). "How Would One Extra Year of High School Affect Academic Performance in University? Evidence from a Unique Policy Change." Working Paper.

[9] Krueger, Alan B. (1999). "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics*, **114,** 497-532.

[10] Lee, Jong-Wha and Robert Barro (2001). "School Quality in a Cross-Section of Countries." *Economica*, **68**, 465-488.

[11] Madalla, G.S. (1986). "Limited-Dependent and Qualitative Variables in Econometrics." Cambridge University Press.

[12] Marcotte, Dave E. (2007). "Schooling and Test Scores: A Mother-Natural Experiment." *Economics of Education Review*, **26,** 629-640.

[13] Marcotte Dave E. and Steven W. Hemelt (2007). "Unscheduled School Closings and Student Performance." IZA DP No. 2923.

[14] Margo, Robert A (1986). "Educational Achievement In Segregated School Systems: The Effects of "Separate-but-Equal." *American Economic Review*, **76,** 794-801.

[15] Miller, Raegen T. and Richard J. Murnane and John B. Willett (2007). "Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One Urban School District." NBER Working Paper No. 13356.

[16] Pischke, Jorn-Steffen (2007). "The Impact of Length of the Schooled Year on Student Performance and Earnings: Evidence From the German Short School Years." *Economic Journal,* **117,** 1216-1242.

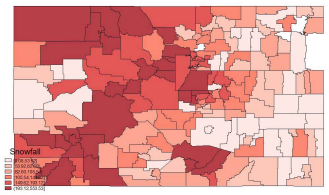
[17] Wößmann, Ludger (2003). "Schooling Resources, Educational Institutions and Student Performance: the International Evidence." *Oxford Bulletin of Economics and Statistics*, **65,** 117-170.

2.6 Appendices

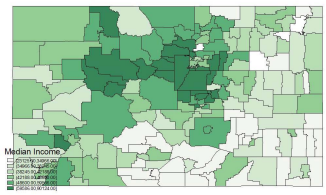
2.6.1 Figures

Appendix Figure 2.1

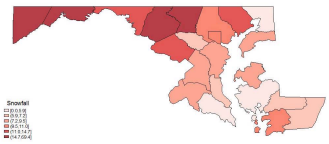
Colorado Mean Yearly Snowfall



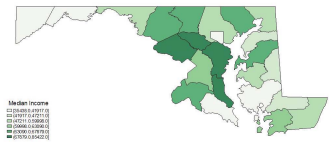
Colorado Median Income



Maryland Mean Yearly Snowfall



Maryland Median Income



2.6.2 Creation of Snowfall Variables

For both Maryland and Colorado, weather data was extracted from the National Climatic Data Center (NCDC) daily surface observations. In Maryland snowfall was taken as the average of snowfall observed in county, as districts and counties are the same. A few counties which did not maintain weather histories and were linked to the closest coop locations. Also days with missing observations were imputed using the closest coops higher and lower in elevation. In Colorado, school districts locations, in longitude and latitude, were extracted from the National Center of Educational Statistics. This was then used to determine the elevation of the school districts. Knowing both the latitude/longitude coordinates of the schools and their elevations, school districts were linked with the two near-

est weather stations higher and lower. Then snowfall was computed the average of these nearby stations. The first stage estimates use the data available from the NCDC as of August 2007.

Part II

Econometrics

Chapter 3

Consistency of Likelihood Ratio Tests for Regime Switching

3.1 Introduction

Tests for regime switching play an important role in analyzing economic data. In the model of Porter (1983) the presence of regimes indicates the exercise of market power, while in the model of Hamilton (1989) regimes offer a compact way to express the dynamic behavior of US output. As multiple regimes are a central feature of these models, accurate testing for regime switching is vital. Yet, as described in more detail below, several features of regime-switching models make accurate inference challenging. Hansen (1992), who revisits the model of Hamilton, develops a limit theory for tests of regime switching, but his approach does not include the boundary of the parameter space. Cho and White (2007), who revisit the model of Porter, include the boundary of the parameter space when developing their limit theory for likelihood ratio tests, but are confronted with the need to specify a parameter space for the coefficients that vary over regimes to obtain critical values. Our focus is to implement the limit theory of Cho and White by addressing two questions. First, to what extent is the power of the likelihood ratio test reduced by lack of knowledge of the regime-varying parameter space? Second, if there is a dramatic reduction in power, can the method of subsampling provide critical values that overcome sensitivity to the parameter space specification and thereby eliminate power losses?

It is well known that the parameters governing switching between regimes are not identified under the null hypothesis of only a single regime. Further,

it is well documented that derivatives of the log-likelihood are identically zero when evaluated under the null hypothesis of only a single regime (Chesher, 1984), regardless of the population value of the regime coefficient. For these reasons likelihood-ratio test statistics are most frequently used to test for regime switching. The asymptotic distribution of the likelihood ratio test depends on the allowable space for the regime-switching probability. Ghosh and Sen (1985), who study regime switching in models that include only an intercept, establish the limit theory when the regime-switching probability is allowed to take the boundary value of 0 or 1. Cho and White extend the result to regime-switching models with additional regressors, and show that allowing the regime-switching probability to take the full range of values in $[0, 1]$ yields a limit distribution that depends on the regime-varying parameter space.

As the limit theory developed by Cho and White is not standard, some form of approximation is needed to obtain critical values. Cho and White present a method of numerical approximation that relies on explict specification of the parameter space. The need to specify an interval for regression coefficients places an additional burden on researchers. As the critical values are obtained by searching over the specified parameter space, researchers must guard against specifying too large a space, as enlarging the space over which the critical value is calculated increases the critical value and reduces power. But researchers must also guard against specifying too small a space. If the specified parameter space is too small,

then the estimates will likely be limited by the boundary of the parameter space, reducing the value of the likelihood ratio test and again reducing power. Thus, for data with wide separation of regimes, which should make regime switching easier to detect, the need to specify intervals can result in almost total loss of power.

In many cases a researcher does not have well defined bounds for these coefficient values. A suitable interval could be obtained by first estimating an unconstrained model, but the resultant test statistic would have to be adjusted to remove the bias arising from the initial estimation. The method of subsampling offers an alternative approximation to obtain critical values.[1] The existence of the limit distribution, as established by Cho and White, ensures that subsampling provides asymptotically valid critical values.[2] As the critical values are obtained by resampling, rather than numeric calculation over the specified parameter space, there is no need to explicitly specify the parameter space. In consequence the power of the likelihood ratio test is not dependent on the researcher's choice of parameter space.

[1]The bootstrap provides an alternative method of resampling to determine critical values. McLachlan (1987) uses the bootstrap to determine significance in the special case of iid data, yet the theory justifying the use of the bootstrap in general tests for regime switching is not well developed. Further, bootstrap approximations may not be accurate when the limit theory depends on the underlying parameter values, such as the regime-switching probability, or when parameters take boundary values (Anderws 2000).

[2]See Andrews and Guggenberger (2007a ,2007c), Linton et al. (2005) and Chernozhukov et al. (2007) for recent applications of subsampling.

3.2 Likelihood-Ratio Tests for Regime Switching

We consider the regime-switching regression

$$Y_t = \theta_j \cdot 1\left(R_t = j-1\right) + X_t'\beta^* + U_t, \tag{3.1}$$

where the latent regime is indexed by the Bernoulli random variable R_t, X_t is a $k \times 1$ vector of observed regressors and $\{U_t\}_{t=1}^n$ is an i.i.d. sequence with $U_t \sim N\left(0, \sigma^{2*}\right)$. The parameters that do not vary over regimes are $\theta_0 = (\beta, \sigma^2)$, so the regime specific intercept is represented as $\theta_1 = \alpha$ and $\theta_2 = \alpha + \gamma$. Further $(\theta_0, \theta_j) \in \Theta_0 \times \Theta_*$ for $j = 1, 2$, where Θ_0 and Θ_* are convex and compact subsets in $\mathbb{R}^2$ and $\mathbb{R}$, respectively. The latent regime is (initially) defined as an i.i.d. sequence with

$$P\left(R_t = 1\right) = \lambda^*. \tag{3.2}$$

Although the model is formulated as a linear regression, the presence of the latent regime leads to maximum likelihood (rather than OLS) as the estimation method. The observation-t value of the log-likelihood is $l_t\left(\lambda, \theta_0, \theta_1, \theta_2\right) = \ln f\left(Y_t | X_t; \lambda, \theta_0, \theta_1, \theta_2\right)$ with

$$f\left(Y_t | X_t; \lambda, \theta_0, \theta_1, \theta_2\right) = (1-\lambda)\frac{e^{-\frac{1}{2\sigma^2}\left(Y_t - \theta_1 - X_t'\beta\right)^2}}{\sqrt{2\pi}\sigma} + \lambda\frac{e^{-\frac{1}{2\sigma^2}\left(Y_t - \theta_2 - X_t'\beta\right)^2}}{\sqrt{2\pi}\sigma}.$$

The unconstrained maximum likelihood estimates are $\left(\hat{\lambda}, \hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2\right) = \arg\max_{\lambda,\theta_0,\theta_1,\theta_2} L_n(\lambda, \theta_0, \theta_1, \theta_2)$ with $L_n(\lambda, \theta_0, \theta_1, \theta_2) = \sum_{t=1}^{n} l_t(\lambda, \theta_0, \theta_1, \theta_2)$. The standard null hypothesis to test for multiple regimes is based on the assumption that $\lambda^* \in (0, 1)$ (so that θ_1^* and θ_2^* are separately identified) from which we form $H_0' : \theta_1^* = \theta_2^* = \theta_*$ and $H_1' : \theta_1^* \neq \theta_2^*$. (Note, these hypotheses could be equivalently expressed as $H_0' : \gamma^* = 0$ and $H_1' : \gamma^* \neq 0$.)

The features of regime-switching models that affect tests of H_0' (as noted in the introduction) are easily seen. The conditional density evaluated under the null

$$f(Y_t|X_t; \lambda, \theta_0, \theta_*, \theta_*) = \frac{e^{-\frac{1}{2\sigma^2}(Y_t - \theta_* - X_t'\beta)^2}}{\sqrt{2\pi}\sigma},$$

does not depend on λ. Hence the parameter governing the behavior of the regime variable (λ^*) is not identified under H_0' and Wald tests, which require limit theory for the unconstrained estimates under the null, are not employed to test for regime switching.

To show that derivatives of the likelihood are identically zero under the null, define the constrained (by H_0') maximum likelihood estimates as $\left(\hat{\theta}_0^c, \hat{\theta}_1^c\right) = \arg\max_{\theta_0,\theta_1} L_n(\lambda, \theta_0, \theta_1, \theta_1)$. The score for θ_1, which is proportional to the score for θ_2, is

$$\bigtriangledown_{\theta_1} L_n(\lambda, \theta_0, \theta_1, \theta_2) = (1 - \lambda) \sum_{t=1}^{n} \frac{e^{-\frac{1}{2\sigma^2}(Y_t - \theta_1 - X_t'\beta)^2}}{\sqrt{2\pi}\sigma f(Y_t|X_t; \lambda, \theta_0, \theta_1, \theta_2)} \frac{(Y_t - \theta_1 - X_t'\beta)}{\sigma^2}.$$

Because the residuals sum to zero, the scores for both θ_1 and θ_2 are numerically zero when evaluated at $\left(\hat{\theta}_0^c, \hat{\theta}_1^c\right)$, regardless of the population values of θ_1 and θ_2. Moreover, $\bigtriangledown_{\theta_2^2} f\left(Y_t|X_t; \lambda, \theta_0, \theta_1, \theta_2\right)$ is proportional to $\left(\frac{(Y_t-\theta_1-X_t'\beta)^2}{\sigma^2} - 1\right)$ and so is also numerically zero when evaluated at the constrained estimates. As a similar argument holds for θ_2, the second derivative of the log-likelihood with respect to either regime variable vanishes at $\left(\hat{\theta}_0^c, \hat{\theta}_1^c\right)$. Hence a test based upon the magnitude of the score evaluated at the null value would never reject the null hypothesis and so the test statistic for regime switching is not based upon the score for the coefficient on the regime variable.[3]

In light of the difficulties associated with tests based directly on the score, tests for regime switching are often based on a log-likelihood ratio. The log-likelihood ratio is

$$LR_n = 2\left[L_n\left(\hat{\lambda}, \hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2\right) - L_n\left(\hat{\theta}_0^c, \hat{\theta}_1^c\right)\right].$$

The LR statistic behaves in unappealing ways if the regime variances differ or if the latent regimes are governed by a Markov process. In the former case the likelihood can be increased by driving the variance of one regime to zero, while in the latter case the variance of the score for the probabilities governing the Markov structure grows geometrically with n. It is for these reasons that we assume that the error variance is equal across regimes in (3.1) and that the regime is generated

[3]For this reason the method of Davies (1977, 1987), which is based on variation in the score function evaluated under the null, is not applicable to regime-switching models.

by an i.i.d. process in (3.2).[4]

One remaining issue is that the empirical size of likelihood ratio tests of H_0' typically exceed the nominal size. To reduce this overrejection, Cho and White expand the null hypothesis to include boundary values for λ^*. The expanded null hypothesis is $H_0 : \theta_1^* = \theta_2^* = \theta_*$; or $\lambda^* = 0, \theta_1^* = \theta_*$; or $\lambda^* = 1, \theta_2^* = \theta_*$. The limit distribution of LR_n under H_0 has two components, one corresponding to $\theta_1^* = \theta_2^* = \theta_*$ and one corresponding to $\lambda^* = 0, \theta_1^* = \theta_*$.[5] Test of the component $\lambda^* = 0, \theta_1^* = \theta_*$ involves the score for λ, which when evaluated under the null is identically zero if $\theta_2 = \theta_*$. In consequence, this component of the limit distribution is discontinuous at θ_* and is obtained by taking the supremum over $\Theta_* \setminus \{\theta_*\}$, which must be compact to ensure the limit result. Cho and White (Theorem 6, p. 1692) establish that under H_0:

$$LR_n \Rightarrow \max \left[\max (0, Z_*)^2 , \sup_{\Theta_* \setminus \{\theta_*\}} \min [0, G(\theta_2)]^2 \right], \tag{3.3}$$

where $\Rightarrow$ indicates weak convergence and Z_* is a standard Gaussian random variable that is correlated with the scaled Gaussian process $G(\theta_2)$. The term $\max (0, Z_*)^2$ is obtained from the convergence of LR_n when $\lambda^* = \frac{1}{2}$. The term $\sup_{\Theta_* \setminus \{\theta_*\}} \min [0, G(\theta_2)]^2$ is obtained from the convergence of LR_n when $\lambda^* = 0$ or 1. The term corresponding to $\left\{\lambda^* \in (0,1) : \lambda^* \neq \frac{1}{2}\right\}$ is bounded by

[4] Levine (1983) and Cho and White (2007) establish that estimators of (3.1) under (3.2) are consistent even if R_t is generated by a Markov process.

[5] The score for the test of the null hypothesis $\lambda^* = 0, \theta_1^* = \theta_*$ against $\lambda^* \in (0,1), \theta_1^* \neq \theta_2^*$ is identical to the score for testing $\lambda^* = 1, \theta_2^* = \theta_*$ by symmetry.

$\sup_{\Theta_* \setminus \{\theta_*\}} \min [0, G(\theta_2)]^2$ and so does not appear in the limit distribution.[6]

The impact of including boundary values for λ^* is to reduce the frequency of rejection of the null hypothesis, which brings the empirical size closer to the nominal size. The decrease in the frequency of rejection of the null hypothesis comes about for two reasons. First, $\sup_{\Theta_* \setminus \{\theta_*\}} \min [0, G(\theta_2)]^2$ exceeds the term for $\left\{\lambda^* \in (0,1) : \lambda^* \neq \frac{1}{2}\right\}$, which would appear in the limit distribution if boundary values for λ were excluded, so the critical values cannot decrease. Second, the supremum in the second term is defined over $\Theta_* \setminus \{\theta_*\}$, which requires specification of Θ_* to obtain critical values. As Θ_* must be compact, specification of Θ_* is equivalent to specifying closed intervals on which θ_1^* and θ_2^* must lie. This, in turn, leads to estimates of the likelihood under these constraints, which can only serve to lower LR_n. In essence, with the limit distribution from (**??**), infrequent large deviations are less likely to be interpreted as regime switching and the likelihood ratio tests is made more robust to outliers.

There are two sources of model dependence in the limit distribution. The first arises from the stochastic process $G(\theta_2)$. The covariance structure of the process depends upon both the specification of (3.1), e.g. the structure of the regressor set, and the specification of the process underlying the latent regime (3.2). A second source of model dependence arises from the parameter space. Consider two possible parameter spaces $\Theta_*^{(1)}$ and $\Theta_*^{(2)}$. If $\Theta_*^{(1)} \subset \Theta_*^{(2)}$, then

[6] The rate of convergence also depends on λ^*. If $\left\{\lambda^* \in (0,1) : \lambda^* \neq \frac{1}{2}\right\}$, the convergence rate is $n^{\frac{1}{6}}$, while if $\lambda^* = \frac{1}{2}$ the convergence rate is $n^{\frac{1}{8}}$. If $\lambda^* \neq \frac{1}{2}$ the data are asymmetric, which speeds the rate of convergence.

$P\left(\sup_{\Theta_*^{(1)}\setminus\{\theta_*\}} \min\left[0, G\left(\theta_2\right)\right]^2 > a\right) \leq P\left(\sup_{\Theta_*^{(2)}\setminus\{\theta_*\}} \min\left[0, G\left(\theta_2\right)\right]^2 > a\right)$ for all a. Hence enlarging Θ_* increases the critical value.

3.3 Critical Values for LR Tests

While (3.3) establishes a limit distribution for the LR statistic, one must still determine the best method to construct critical values. We detail how to construct critical values via the approximation method in Cho and White and via subsampling. Of particular importance; the approximation method of Cho and White requires calculations that are specific to each model, while the method of subsampling does not.

The Cho-White approximation method first requires that one calculate $E\left[Z_* G\left(\theta_2\right)\right]$ and $E\left[G\left(\theta_2\right) G\left(\theta_2'\right)\right]$ for distinct values θ_2 and θ_2'. For each value of θ_2, the stochastic process $G\left(\theta_2\right) \sim N\left(0,1\right)$, but the covariance structure of $\left(Z_*, G\left(\theta_2\right)\right)$ depends on the specification of (3.1). We focus on the autoregressive model, which captures the dynamic behavior modeled in Hamilton

$$Y_t = \theta_j \cdot 1\left(R_t = j-1\right) + .5Y_{t-1} + U_t, \tag{3.4}$$

where $\sigma^{2*} = 1$. For the autoregressive model

$$\begin{aligned} E\left[Z_* G\left(\theta_2\right)\right] &= \frac{\theta_2^4}{e^{\theta_2^2} - 1 - \theta_2^2 - \frac{\theta_2^4}{2}} \\ E\left[G\left(\theta_2\right) G\left(\theta_2'\right)\right] &= \frac{e^{\theta_2\theta_2'} - 1 - \theta_2\theta_2' - \frac{\left(\theta_2\theta_2'\right)^2}{2}}{\left(e^{\theta_2^2} - 1 - \theta_2^2 - \frac{\theta_2^4}{2}\right)^{\frac{1}{2}} \left(e^{\left(\theta_2'\right)^2} - 1 - \left(\theta_2'\right)^2 - \frac{\left(\theta_2'\right)^4}{2}\right)^{\frac{1}{2}}}. \end{aligned}$$

The next step is to construct a Gaussian process that has the same covariance structure of $(Z_*, G(\theta_2))$ and, therefore, has a distribution that is identical to the distribution of $G(\theta_2)$. For $\{Y_i\}$ a sequence of independent $N(0,1)$ random variables, the constructed Gaussian process for the autoregressive model is

$$f\left(\theta_2\right) = \sum_{i=3}^{\infty} \frac{\theta_2^i Y_i}{\left\{i!\left[\exp\left(\theta_2^2\right) - 1 - \theta_2^2 - \theta_2^4/2\right]\right\}^{\frac{1}{2}}}. \tag{3.5}$$

For the constructed Gaussian process, simulated critical values are based on the sum truncated at $i = 150$. For the autoregressive model

$$\tilde{f}\left(\theta_2\right) = \sum_{i=3}^{150} \frac{\theta_2^i Y_i}{\left\{i!\left[\exp\left(\theta_2^2\right) - 1 - \theta_2^2 - \theta_2^4/2\right]\right\}^{\frac{1}{2}}}.$$

We then evaluate $\tilde{f}(\theta_2)$ for each value in $grid(\Theta_*)$, where $grid(\Theta_*)$ denotes a grid of Θ_* with mesh size 0.01. The autoregressive model critical value is

$$\max\left\{\left[\max\left(0, Y_4\right)\right]^2, \sup_{\theta_2 \in grid(\Theta_*)} \left\{\min\left[0, \tilde{f}\left(\theta_2\right)\right]\right\}^2\right\}. \tag{3.6}$$

Note, Y_4 appears in the first term to generate the correct covariance between Z_* and $G(\theta_2)$.

The subsampling method to obtain critical values does not depend on the specification of (3.1). To implement the method, we first construct subsamples that consist of (overlapping) blocks of the data. $\left\{(Y_s, X_s'), \ldots, \left(Y_{s+b-1}, X_{s+b-1}'\right)\right\}_{s=1}^{n-(b-1)}$, where $b = n^{\frac{1}{2}}$ is the subsample length. For each subsample, indexed as (b, s) we estimate the likelihood ratio

$$LR_{b,s} = 2\left[L_{b,s}\left(\hat{\lambda}_{b,s}, \hat{\theta}_{0,b,s}, \hat{\theta}_{1,b,s}, \hat{\theta}_{2,b,s}\right) - L_{b,s}\left(\hat{\theta}^c_{0,b,s}, \hat{\theta}^c_{1,b,s}\right)\right],$$

with $\{LR_{b,s}\}_{s=1}^{n-(b-1)}$ the sequence of likelihood ratios estimated from each of the $n-(b-1)$ subsamples. Next order the estimated likelihood ratios from smallest to largest, yielding the order statistics $\left\{LR_{b,(s)}\right\}$. For a test with size 5 percent, the critical value is formed from $LR_{b,(r)}$, where r is the first integer at least as large as $(.95)(n-b+1)$. One then rejects H_{0c} if $LR_n > \left(\frac{b}{n}\right)^{\tau} LR_{b,(r)}$, where the scaling factor τ is the rate of convergence of LR_n under H_{0c}. As the regime-switching specifications under study generate stationary time-series data, the limit distribution (3.3) together with Theorem 3.2.1 in Politis, Romano and Wolf (1998) implies that the test based on subsampling critical values is asymptotically valid.

3.4 Impact of Parameter Space Specification

Specification of the parameter space has two impacts on the likelihood ratio statistic with critical values from the Cho and White approximation, LR_{CW}. First, as the estimates of θ_2 (and θ_1) in the unconstrained model cannot lie outside Θ_*, reducing Θ_* increases the frequency with which $\hat{\theta}_2$ equals a boundary value of Θ_*. In consequence $\hat{L}$ is reduced thereby reducing LR_{CW}. To overcome the boundary value problem one can enlarge Θ_*. Yet enlarging Θ_* results in the second impact: as seen in the previous section enlarging Θ_* raises the critical value. To understand the impact of these two competing effects, we focus on the autoregressive model (3.4) for which the approximation formulae are presented in Section 3. Recall that Θ_* is the allowable parameter space for θ_2^*, so Θ_* is simply an interval for θ_2^*. In practice, researchers may have limited information about θ_1^* and θ_2^*, and so little guide to correct specification of Θ_*. As theory requires that Θ_* contain θ_2^*, one possible solution is to select a large interval for Θ_*. Yet (3.6) indicates that critical values tend to increase as Θ_* increases, which reduces power and so imposes a cost when enlarging Θ_*. To better understand the impact of the interplay between θ_2^* and Θ_*, we study the performance of the Cho and White critical values for three scenarios: θ_2^* interior to Θ_*, θ_2^* on the boundary of Θ_* and θ_2^* outside Θ_*. To form these scenarios we consider values of $\theta_2^* \in \{0, 0.4, 0.8, \ldots, 4.0\}$. For each value of θ_2^* we construct the critical value

(3.6) for four specifications of Θ_*:

$$\Theta_* = \{-j, j\} \text{ for } j = 1, 2, 3, 4.$$

This produces cases in which Θ_* is both too small (not containing θ_2^*) and too large (Θ_* could shrink considerably and still contain θ_2^*). We employ symmetric intervals, rather than intervals of the form $[0, j]$, to avoid boundary issues for the estimate of θ_1, which is frequently negative.

As both skewness and kurtosis reflect evidence of regime switching, the power also depends upon the probability of regime switching, λ^*. Because the likelihood function is symmetric in $(\lambda^*, \theta_1^*, \theta_2^*)$ and $(1 - \lambda^*, \theta_2^*, \theta_1^*)$, we need only consider values of λ^* on one half of the unit interval. We select $\lambda^* \in [0, 0.5]$ so that the unconditional mean of the process generally lies within Θ_*, to reduce the likelihood that estimates are constrained by boundary values.

Specification of the interval has two impacts on the size of LR tests. First, as noted above, specification of a larger interval for Θ_* will generally increase the critical values, thereby lowering the empirical size and power of the test. A second, countervailing, impact is the effect of the parameter space on $L_n\left(\hat{\lambda}, \hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2\right)$. Because the estimated value of θ_2 must be contained in Θ_*, specification of a smaller interval for Θ_* is more likely to force $\hat{\theta}_2$ to be constrained on the boundary of Θ_*, which lowers LR_n. Intuitively, constraining $\hat{\theta}_2$ to be on the boundary of

Table 3.1: Coefficient Interval Specification: Empirical Test Size

Θ_*	$[-1,1]$	$[-2,2]$	$[-3,3]$	$[-4,4]$
Size	5.6%	7.1%	5.1%	4.6%
$\hat{\theta}_2$ at boundary	22%	<1%	0%	0%
Power	0%	26.3%	28.6%	26.9%
$\hat{\theta}_2$ at boundary	91%	52%	1%	<1%

Θ_* raises $\hat{\sigma}_n^2$, which reduces the evidence of regime switching. As $\hat{\theta}_2$ is less likely to be constrained to a boundary value as Θ_* takes larger intervals, LR_n generally increases with specification of a larger interval, thereby raising the empirical size and power of the test.

In the upper panel of Table 3.1 we report the empirical test size, for a nominal test size of 5 percent, together with the frequency with which the estimate attains the boundary value.[7]

The upper panel clearly reveals the impact of Θ_*. The effect of enlarging Θ_* from $[-1,1]$ to $[-2,2]$ reduces the fraction of estimates limited by the boundary value from 22 percent to less than 1 percent. The removal of the boundary constraint raises the value of LR by more than the accompanying increase in the critical value and the size increases. Further enlargement of Θ_* only serves to increase the critical value, and the size monotonically declines. The lower panel contains the test power for $\theta_1^* = 0, \theta_2^* = 2.0$. Again, the impact of Θ_* is

[7]Reported estimates are obtained from the EM algorithm; estimates obtained via grid search are similar. For each simulated data set, the initial values of the estimates under the null are the sample mean and variance of Y (for the intercept and σ^2, respectively). Under the alternative, the initial values also include $\lambda = .5$ and $\theta_1 = \bar{Y}_n - c$, $\theta_2 = \bar{Y}_n + c$, where c is the value from $\{.2, .4, \ldots, 2.0\}$ that maximizes the likelihood.

pronounced. When Θ_* does not contain θ_2^*, virtually all of the estimated values are limited by the boundary value and the power falls below the size. Enlarging Θ_* from $[-1, 1]$ to $[-3, 3]$ again greatly reduces the fraction of estimated values limited by the boundary value, and the power increases markedly. Further enlargement of Θ_* serves only to increase the critical value and the power declines.

The interplay between specification of Θ_* and the power of the test as θ_2^* increases, is made clearer in Figure 3.1. In each panel, we see that the power (displayed on the left scale) increases with θ_2^* until θ_2^* reaches the boundary of Θ_*. At this point, further increases in θ_2^* leave $\hat{\theta}_2$ constrained at the boundary value, thereby increasing the estimated variance $\hat{\sigma}^2$ (displayed on the right scale). Larger values of λ^* lead both to an increased chance of a constrained estimate (as the unconditional mean of the process is larger) and to larger weight given to these residuals, thereby increasing $\hat{\sigma}^2$. In the upper panel, at $\theta_2^* = 3.0$, the estimated variance is more than twice as large as the actual variance and the empirical power is effectively zero. To guard against the severe consequences of specifying a coefficient interval that is too small, researchers should err on the side of larger Θ_*. If the specified interval is smaller, the LR test based on the approximate critical values suffers size distortion. If the interval is larger, the LR test suffers power loss.

In Table 3.2 we study how altering the degree of skewness impacts the results. As in Table 3.1, enlarging Θ_* leads to power gains only in as much as the boundary

Figure 3.1: Impact of Parameter Space on Power, $\lambda^* = .3$

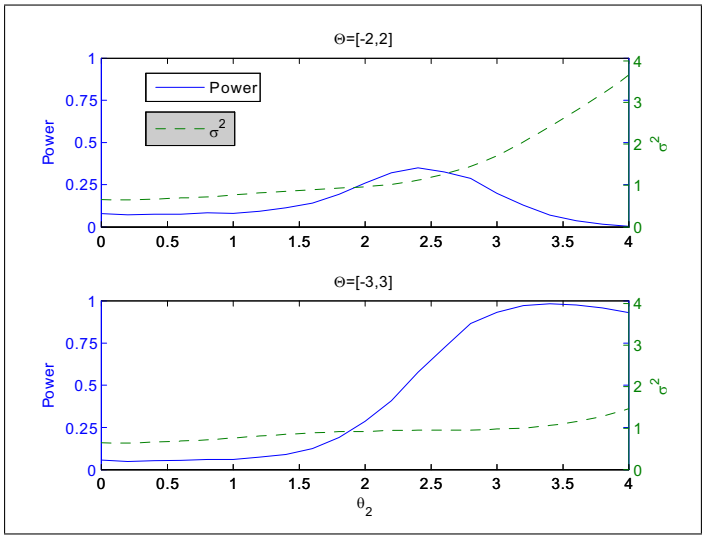


Table 3.2: Altering Skewness

Θ_*		$[-1,1]$	$[-2,2]$	$[-3,3]$	$[-4,4]$
$\lambda = .1$	Power	1%	23.8%	22.6%	21.7%
$\hat{\theta}_2$ at boundary		62%	25%	3%	<1%
$\lambda = .5$	Power	0%	18.7%	23.2%	22.2%
$\hat{\theta}_2$ at boundary		93%	58%	<1%	0%

limit on $\hat{\theta}_2$ is relaxed. Increasing asymmetry, as λ changes from .3 to .1, reduces the likelihood of regime switching and the associated power. When asymmetry vanishes, regime switching becomes frequent, but with little tendency to remain in a given regime it is difficult to identify regime switching. Recall that for $\lambda = .5$ the third derivative of the log-likelihood with respect to θ_2 also vanishes and estimation of θ_2 is considerably more difficult (the rate of convergence slows from $n^{\frac{1}{6}}$ to $n^{\frac{1}{8}}$).

3.5 Performance of Subsample Critical Values

Previous findings from Andrews and Guggenberger (2007 a, b, and c) suggest the lack of uniform convergence in subsampling can cause subsample-based statistics to exhibit differences in behavior over sample sizes or parameter values. To determine the accuracy of subsample critical values, we study the empirical size and power of LR tests. We compare the LR test based on subsample critical values with three alternatives: a test based on excess skewness and kurtosis; a modified version of the $C(\alpha)$ test of Neyman and the LR test based on critical values from the Cho and White approximation. Because the critical values from the Cho and White approximation depend on specification of Θ_*, we first study how Θ_* affects the size and power of the LR test that uses these critical values.

The previous section reveals the severe power loss for LR tests that can arise with critical values calculated via the Cho and White approximation method. The

likelihood ratio statistic with subsample critical values, LR_{Sub}, does not suffer from the same power loss. The reason; construction of the subsample critical value does not require explicit specification of Θ_*. Hence the unconstrained estimates are not limited by the boundary of the parameter space, opening the way for large increases in power with size controlled. To determine the performance of subsample critical values, we compare the empirical size and power of LR_{Sub} with two moment-based tests. Because either excess skewness or kurtosis (in relation to a normal distribution) can reveal evidence of regime switching, the two benchmark statistics we study are functions of the sample skewness, s_n, and kurtosis, k_n.[8] The Jarque-Bera statistic is

$$JB = n\left(\frac{s_n^2}{6} + \frac{(k_n - 3)^2}{24}\right),$$

which follows a chi-square distribution of with 2 degrees freedom under the null. The second benchmark statistic is based on the $C(\alpha)$ statistic of Neyman. Cho and White modify the $C(\alpha)$ statistic to account for the zero second derivative of the log-likelihood, yielding

$$C(\alpha) = n \max\left[\frac{s_n^2}{6}, \min\left[0, \frac{k_n - 3}{24^{1/2}}\right]^2\right].$$

The limit distribution of the $C(\alpha)$ statistic is $\max\left[Z_1^2, \min\left[0, Z_2\right]^2\right]$, where Z_1 and

[8] These sample quantities are formed from the residuals generated by the null estimates $\left(\hat{\theta}_0^c, \hat{\theta}_1^c\right)$.

Table 3.3: Autoregressive Model

Test	LR_{Sub}	$C(\alpha)$	JB	LR_{CW}
Size	3.6%	3.0%	3.8%	5.1% to 7.1%
Power	23.5%	10.0%	6.2%	0% to 28.6%
$\lambda = .1$	17.6%	23.9%	23.6%	1% to 23.8%
$\lambda = .5$	18.2%	1.2%	0.04%	0% to 23.2%

Z_2 are independent Gaussian random variables.

We study three models: the autoregressive model (3.4); a mixture model and a simultaneous equations model. Given the complexity of the models, the sample size we study is $n = 100$. The scaling factor is $\left(\frac{b}{n}\right)^{\frac{1}{6}}$, which accords with the limit theory for all $\lambda^* \neq .5$.

3.5.1 Autoregressive Model

To compare the performance of the three test statistics, we first return to the autoregressive model (3.4). In Table 3.3, we present the empirical test size, for a nominal size is 5 percent, and power for the autoregressive model. On the second row, we present test power for $\theta_1 = 0 \quad \theta_2 = 2.0 \quad \lambda = .3$. The lower panel contains test power for the same values of θ_1 and θ_2, but for different values of the regime probability λ.[9] The final column presents the range of values for size and power of the LR test statistic with the critical value obtained by the Cho and White approximation.

[9]Unless otherwise noted, all simulations are based on 3000 Monte Carlo replications.

The three tests are all conservative with the subsample LR as close to the nominal size as the $C(\alpha)$ test. Two important points emerge. First, the LR test statistic with subsampled critical values is far more powerful than the two statistics based on skewness and kurtosis. Second, LR_{Sub} achieves more than 80 percent of the maximum power of LR_{CW} without the potential great loss of power from misspecification of the parameter space. Clearly the power is not a monotone function of λ^* for given θ_2^*, for $\lambda^* = 0$ there is no power to detect departures from the null.

In Figure 3.2, we present the power curves for the full range of θ_2^* under the alternative hypothesis. Separation of the power curves is most pronounced for the more difficult testing problems in which $\lambda^* \in \{.3, .5\}$. Also for any of the test statistics, notable separation between the two states is needed to achieve moderate power.

3.5.2 Mixture Model

To determine the impact of the autoregressive coefficient on the power to detect regime switching, we also study the mixture model that forms the basis of much statistical analysis of testing for regime switching

$$Y_t = \theta_j \cdot 1(R_t = j - 1) + 4X_t + U_t, \tag{3.7}$$

Figure 3.2: Auto-Regressive Model Power Curves

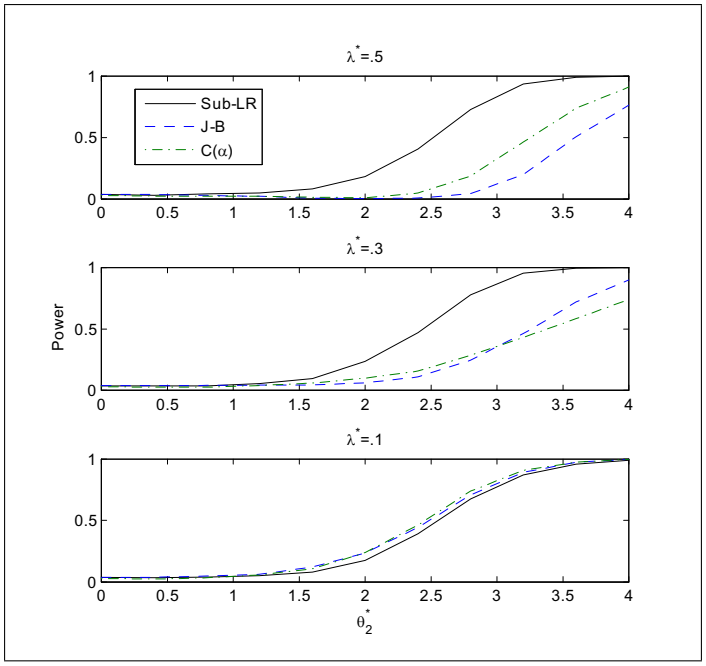


Table 3.4: Mixture Model

Test	LR_{Sub}	$C(\alpha)$	JB
Size	2.5%	3.5%	4.0%
Power	16.7%	8.5%	5.7%
$\lambda = .1$	13.7%	23.7%	22.6%
$\lambda = .5$	16.2%	1.5%	0.005%

with $\theta_0^* = \sigma^{2*} = 1$. As the mixture model has no dependence, there is less potential separation between regimes, making regimes harder to detect. In the upper panel of Table 3.4, for which $\lambda = .3$, we present the empirical test size on the second row (again, nominal size is 5 percent). On the third row, we present test power for $\theta_1 = 0 \quad \theta_2 = 2.0$.

The power of LR_{Sub} dominates the power of the JB and $C(\alpha)$ statistics for most

Table 3.5: Mixture Model Size-Adjusted Power

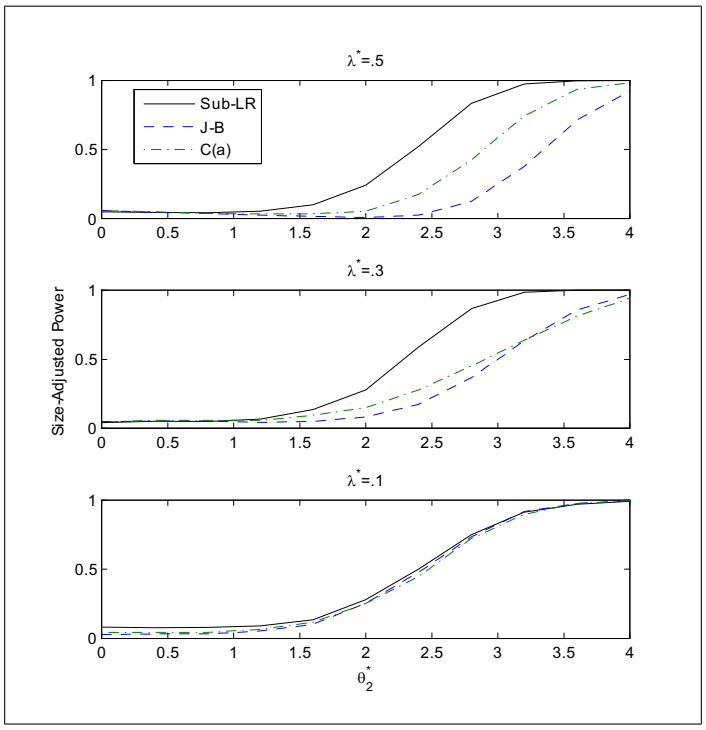
Test	LR_{Sub}	$C(\alpha)$	JB
$\lambda = .3$	27.6%	14.9%	8.2%
$\lambda = .1$	23.6%	32.8%	28.1%
$\lambda = .5$	24.1%	5.3%	0.8%

values of λ, although as λ approaches 0 or 1 and skewness increases, the power of the JB and $C(\alpha)$ tests. As conjectured, the decline in dependence reduces power.

In the lower panel of Table 3.4, we study how altering skewness impacts the power of the test statistics. As λ declines one regime becomes increasingly more likely, which increases skewness and kurtosis. As the JB and $C(\alpha)$ test statistics are sensitive to changes in both these measures, their performance is nearly equal to LR_{Sub} when $\lambda = .1$. For $\lambda = .5$, however, skewness vanishes as the underlying distribution is symmetric around 0. Further, while the kurtosis falls below 3, the absence of skewness reduces the power of the JB and $C(\alpha)$ test statistics to such an extent that power falls below nominal size. Indeed, it appears that the variation in the measure of skewness under the null exceeds the variation under the alternative, which drives the power of these statistics below the nominal size.

In Table 3.5 we report the power when the statistics for $\theta_1 = 0 \quad \theta_2 = 2.0$. We see that the power gains of the LR_{Sub} statistic largely remain when correcting for size distortion. Curves displaying the power for $\theta_2^* \in \{0.4, 0.8, \ldots, 4.0\}$ are contained in Figure 3.3.

Figure 3.3: Mixture Model Power Curves



3.5.3 Simultaneous Equations Model

The third specification includes a second endogenous variable to capture the potential collusion of Porter's model in an industry that also produces a good competitively (as in the petroleum markets studied in Griffin et al. 2006)

$$\begin{aligned} Y_{1t} &= \theta_j \cdot 1\left(R_t = j-1\right) + X_{1t} + .5Y_{2t} + U_{1t} \\ Y_{2t} &= X_{2t} - .5Y_{1t} + U_{2t}, \end{aligned} \tag{3.8}$$

where $\{X_{1t}\}$ and $\{X_{2t}\}$ are mutually independent sequences of i.i.d. $N(0,4)$ random variables and $\{U_{1t}\}$ and $\{U_{2t}\}$ are mutually independent sequences of i.i.d. $N(0,1)$ random variables. The larger variance of the regressors ensures that the model is empirically well identified. The increased number of parameters for the simultaneous equations model leads to a larger subsample length to ensure adequate degrees of freedom. As there are 10 parameters to estimate under the alternative hypothesis, note that we must estimate an intercept for the second equation even though the population value of this coefficient is zero and that the two regression errors have distinct variances, $b = n^{\frac{1}{2}} + 10$. In Table 3.6, we present the empirical test size on the first row (again, nominal size is 5 percent). On the second row, we present test power for $\theta_1 = 0 \quad \theta_2 = 2.0 \quad \lambda = .5$.

Although power is low, the LR_{Sub} test statistic delivers a considerable power gain over the other test statistics, except as λ gets quite close to either 1 or 0, in which

Table 3.6: Simultaneous Equations Model

Test	LR_{Sub}	$C(\alpha)$	JB
Size	3.2%	2.8%	2.8%
Power	18.4%	8.4%	6.8%
$\lambda = .1$	14.6%	16.6%	18.4%
$\lambda = .5$	17.0%	2.4%	0.8%

case the increased skewness levels provide more power for those tests.

3.6 Conclusions

Two principal conclusions emerge. To obtain a critical value for the LR test statistic, the method of subsampling provides an attractive alternative to numeric approximation of the limit distribution. Second, the LR test statistic with subsampled critical values is more powerful than competing statistics based on skewness and kurtosis.

The method of subsampling has two important features. First, there is no need to devise an approximation function for each model. Second, in selecting the parameter space, there is no cost in forgone power from allowing Θ_* to be a large interval. Indeed, for the autoregressive model under study, the LR test statistic with subsample critical values achieves more than 80 percent of the maximum power of the approximation critical values, without the risk of severe power loss that can occur with critical values obtained from numeric approximation.. The subsample size could vary with the dependence in the process and further research

could be helpful here. Finally, while it may be appealing to pre-test and so define Θ_* based on characteristics of the data, the limit theory has not been established for such a case.

In future work we hope to explore the possibility of estimating the scaling factor via multiple block sizes. As the rate of convergence, and hence the scaling factor, depend on the unknown parameter λ, we will employ multiple block sizes to estimate the factor. Let the block sizes be $b_1 = n^{\frac{1}{2}}$ and $b_2 = n^{3/4}$. Politis, Romano and Wolf show (p. 179) that if $\tau(b,n) = \left(\frac{b}{n}\right)^d$, then d can be consistently estimated from

$$\left[\log\left(\frac{b_1}{b_2}\right)\right]^{-1}\left(\log LR_{b_2,(t_i)} - \log LR_{b_1,(t_i)}\right) = d + o_P\left(\left[\log\left(\frac{b_1}{b_2}\right)\right]^{-1}\right), \quad (3.9)$$

where for $i = 1,\ldots,10$, $t_i = .1i(n-b+1)$ indicates the i^{th} decile of the ordered ratios. For a test with size α, we reject H_{0c} if $LR_n > \hat{\tau}(b,n)\, LR_{b,(s)}$, where $\hat{\tau}(b,n) = \left(\frac{b}{n}\right)^{\hat{d}}$. The null hypothesis H_{0c} is then rejected if $LR_n > \hat{\tau}(b,n)\, LR_{b,(s)}$; a procedure that is justified by Politis, Romano and Wolf (Lemma 8.2.1, p. 178). Note, the ability to distinguish between convergence rates of $n^{\frac{1}{6}}$ and $n^{\frac{1}{8}}$ may require large samples.

One could also determine the power to detect Markov switching. To determine the power of the test to detect regimes with Markov switching, one could also

consider specifications in which the latent regime is governed by

$$P\left(R_t = 1 | R_{t-1} = 1\right) = \lambda_{22},$$

while $P\left(R_t = 0 | R_{t-1} = 0\right) = \lambda_{11}$.

Lastly, data driven methods to choose coeficient intervals may be another method to avoid the potential losses when a researcher faces an unknown true alternative. However, the current limit theory does not consider such data driven methods. Further advances incoporate the effect of pre-estimation to select an "optimal" parameter space are necessary to avoid size inflation. If a distribution for such a modified limit theory can be shown to exist, a fixed critical value may provide improved power if size can be controlled.

Bibliography

[1] Andrews, D. and P. Guggenberger (2007). "The Limit of Finite Sample Size: A Problem with Subsampling." *Cowles Foundation Discussion Papers.*

[2] Andrews, D. and P. Guggenberger (2007). "Hybrid and Size-Corrected Subsample Methods." *Cowles Foundation Discussion Papers.*

[3] Andrews, D. and P. Guggenberger (2007). "Applications of Subsampling, Hybrid, and Size-Correction Methods." *Cowles Foundation Discussion Papers.*

[4] Chesher, A. (1984). "Testing for Neglected Heterogeneity." *Econometrica,* **52**, 865-872.

[5] Chernozhukov V., H. Hong and E. Tamer, "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica*, 2007, **75**, 1243-1284.

[6] Cho, J. and H. White (2007). "Testing for Regime-Switching." *Econometrica*, **75**, 1671-1720.

[7] Cho, J. and H. White (2008). "Testing for Unobserved Heterogeneity in Exponential and Weibull Duration Models." *Korea University Discussion Paper.*

[8] Davies, R. (1977). "Hypothesis testing when a nuisance parameter is present only under the alternative." *Biometrika*, **64**, 247-254.

[9] Davies, R. (1987). "Hypothesis testing when a nuisance parameter is present only under the alternative." *Biometrika*, **74**, 33-43.

[10] Dufour, J. (2006). "Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics." *Journal of Econometrics*, **133**, 443-477.

[11] Griffin, J., C. Kolstad and D. Steigerwald (2009). "Estimating Market Power with Multiple Product Markets." *UC Santa Barbara Discussion Papers.*

[12] Hamilton, J. (1989). "A New Approach to the Economic Analysis of Non-Stationary Data." *Econometrica,* **57**, 357-384.

[13] Hansen, B. (1991). "Inference When a Nuisance Parameter is not Identified Under the Null Hypothesis." *Manuscript.*

[14] Hansen, B. (1992). "The Likelihood Ratio Statistic Under Nonstandard Conditions: Testing the Markov-Switching Model of GNP." *Journal of Applied Econometrics,* **7**, S61-S82.

[15] Hansen, B. (1996). "Inference When a Nuisance Parameter is not Identified Under the Null Hypothesis." *Econometrica*, **64**, 413-430.

[16] Levine, D. (1983). "A Remark on Serial Correlation in Maximum Likelihood." *Journal of Econometrics* **23**, 337-342.

[17] Lindsay, B. (1995). *Mixture Models: Theory, Geometry, and Applications*, Hayward Institute for Mathematics and Statistics.

[18] Linton, O., E. Maasoumi and Y. Whang (2005). "Consistent Testing for Stochastic Dominance under General Sampling Schemes." *Review of Economic Studies*, **72**, 735-765.

[19] McLachlan, G. (1987). "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture." *Applied Statistics*, **36**, 318-324.

[20] Politis, D. and J. Romano (1994). "Large Sample Confidence Regions Based On Subsamples under Minimal Assumptions." *The Annals of Statistics*, **22**, 2031-2050.

[21] Politis, D, J. Romano and M. Wolf (1998). *Subsampling*, Springer-Verlag: New York.