Stat 133 Final Report: Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement

Lauren Hanlon, Nadav Tadelis, Daniel Saedi

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Version:  1.35
## Date:     2015-04-25
## Author:   Philip Leifeld (University of Konstanz)
##
## Please cite the JSS article in your publications -- see citation("texreg").
## Loading required package: car
## Loading required package: survival
## Warning: package 'ggplot2' was built under R version 3.2.4
```

Introduction

The goal of this project is twofold. First, to expand our knowledge of statistical tools in R and their potential applications to academic data. Second, to check the reproducibility of Joshua D. Angrist and Victor Lavy's 1999 paper 'Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement' by re-running the paper's regression models in R (compared to thier original analysis which was done in Stata). This is possible because the data used in the paper is publibly posted along with some of the Stata .do files (which were used as reference for the data cleaning) and the models used in the paper are explained in detail.

First Contact

The raw data that Angrist and Lavy post online are posted in .dta format, which R cannot read. In order to make the raw data readable we use the package Rio to import the .dta data into a .cvs file.

Data Cleaning

We clean the data by removing or adjusting all impossible observations (scores over 100%, or schools with no test results), this was done in accordance to Angrist and Lavy's data cleaning. Additionally, we remove the observations with NA's in our regressor variables (the variables we use as regressors later in the project);

this is necessary because in order to calculate clustered standard errors (see `robust.se` function in the data analysis R script in the code folder of the project) the variables need to be equal in length, otherwise `tapply` returns an error.

We then saved the clean data to our data folder in .csv format.

Describing the data

Our cleaned data frames for 5th and 4th graders have 50 columns and 2019 and 2049 rows, respectively. The rows correspond to the number of schools that were observed for this study. Column names correspond to variables that were measured. Variable definitions are as follows (as defined by Angrist and Lavy): Class size = number of students in the spring. Enrollment = September grade enrollment, Percent disadvantaged = percent of studetns in the school from "disadvantaged backgrounds," Reading size = number of students who took the reading test, Math size = number of students who took the math test, Average verbal = average composite reading score in the class, Average math = average composite math score in the class.

We created a discontinuity sample which includes enrollment 36-45, 76-85 and 116-124. This data is used throughout the project.

TABLE I

Unweighted Descriptive Statistics

Table I gives an insight into general statistics about the data. We created tables for the full samples of 5th and 4th grade as well as tables for the discontinuous samples of 5th and 4th grade. The tables are as follows: Full sample 5th grade (2019 classes, 1002 schools, tested in 1991) Full sample 4th grade (2049 classes, 1013 schools, tested in 1991) +/- 5 Discontinuity sample 5th grade (471 classes, 224 schools) +/- 5 Discontinuity sample 4th grade (415 classes, 195 schools)

The table gives data on class size, enrollment, percent disadvantaged, reading size, math size, average verbal score and average math score. The table displays the mean, standard deviation, min, max and quantiles for each variable.

A. Full Sample

Table 1: Unweighted Descriptive Statistics for 5th Grade: 2019 classes, 0 schools, tested in 1991

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Class size	29.935	6.546	8	26	31	35	44
Enrollment	77.742	38.811	8	50	72	100	226
Percent disadvantaged	14.113	13.554	0	4	10	20	76
Reading size	27.323	6.582	5	23	28	32	41
Math size	27.720	6.642	0	23	28	33	41
Average verbal	74.386	7.684	34.800	69.855	75.407	79.843	93.860
Average math	67.259	9.712	0.000	61.100	67.780	74.085	93.930

B. +/- 5 Discontinuity Sample

In the discontinuity sample we narrowed down the sample sizes significantly.

Table 2: Unweighted Descriptive Statistics for 4th Grade: 2049 classes, 0 schools, tested in 1991

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Class size	30.277	6.327	7	26	31	35	44
Enrollment	78.302	37.712	8	51	74	101	225
Percent disadvantaged	13.878	13.385	0	4	9	19	76
Reading size	27.667	6.540	5	24	28	32	69
Math size	28.080	6.548	5	24	29	33	69
Average verbal	72.489	7.991	24.290	67.670	73.330	78.210	94.900
Average math	68.864	8.768	21.410	63.590	69.330	74.970	94.091

Table 3: Unweighted Descriptive Statistics for 5th Grade: 471 classes, 0 schools, tested in 1991

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Class size	30.820	7.418	14	24	31	38	44
Enrollment	76.382	29.503	36	43	79	85	125
Percent disadvantaged	13.597	13.173	0	4	10	17	76
Reading size	28.138	7.333	9	22	28	35	41
Math size	28.541	7.442	9	22	28	35	41
Average verbal	74.491	8.181	34.800	69.655	75.640	80.455	93.810
Average math	67.032	10.203	27.690	60.880	67.360	73.728	92.690

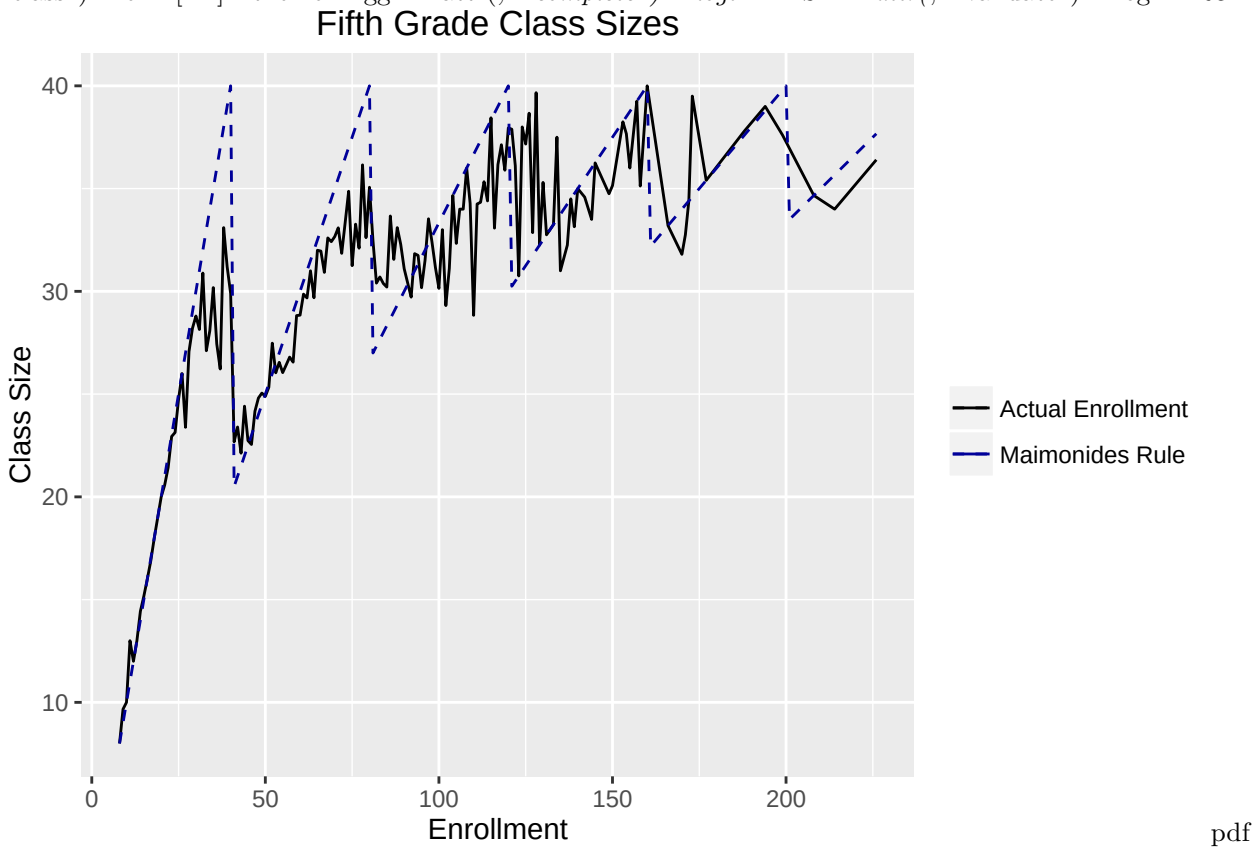
Table 4: Unweighted Descriptive Statistics for 4th Grade: 415 classes, 0 schools, tested in 1991

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Class size	31.149	7.234	12	25	32	38	44
Enrollment	78.547	29.966	36	43	80	116	125
Percent disadvantaged	12.889	12.335	0	4	9	17.8	60
Reading size	28.313	7.704	9	22	28	35	69
Math size	28.749	7.688	11	23	29	35	69
Average verbal	72.468	7.832	46.387	67.000	73.320	78.315	94.900
Average math	68.674	9.085	37.250	62.655	69.340	75.410	92.670

Class Size in 1991 in Initial Enrollment Count, Actual Average Size and as Predicted by Maimonides' Rule

The graph we created compares average class size per enrollment category to expected enrollment according to Maimonides Rule. It shows that the class size behavior (with respect to enrollment) does approximately follow the discontinuous model of the Maimonides' Rule

List of 1 $ plot.title:List of 10 ..$ family : NULL ..$ face : chr "bold" ..$ colour : NULL ..$ size : num 12 ..$ hjust : NULL ..$ vjust : NULL ..$ angle : NULL ..$ lineheight: NULL ..$ margin : NULL ..$ debug : NULL ..- attr(*, "class")= chr [1:2] "element_text" "element" - attr(*, "class")= chr [1:2] "theme" "gg" - attr(*, "complete")= logi FALSE - attr(*, "validate")= logi TRUE

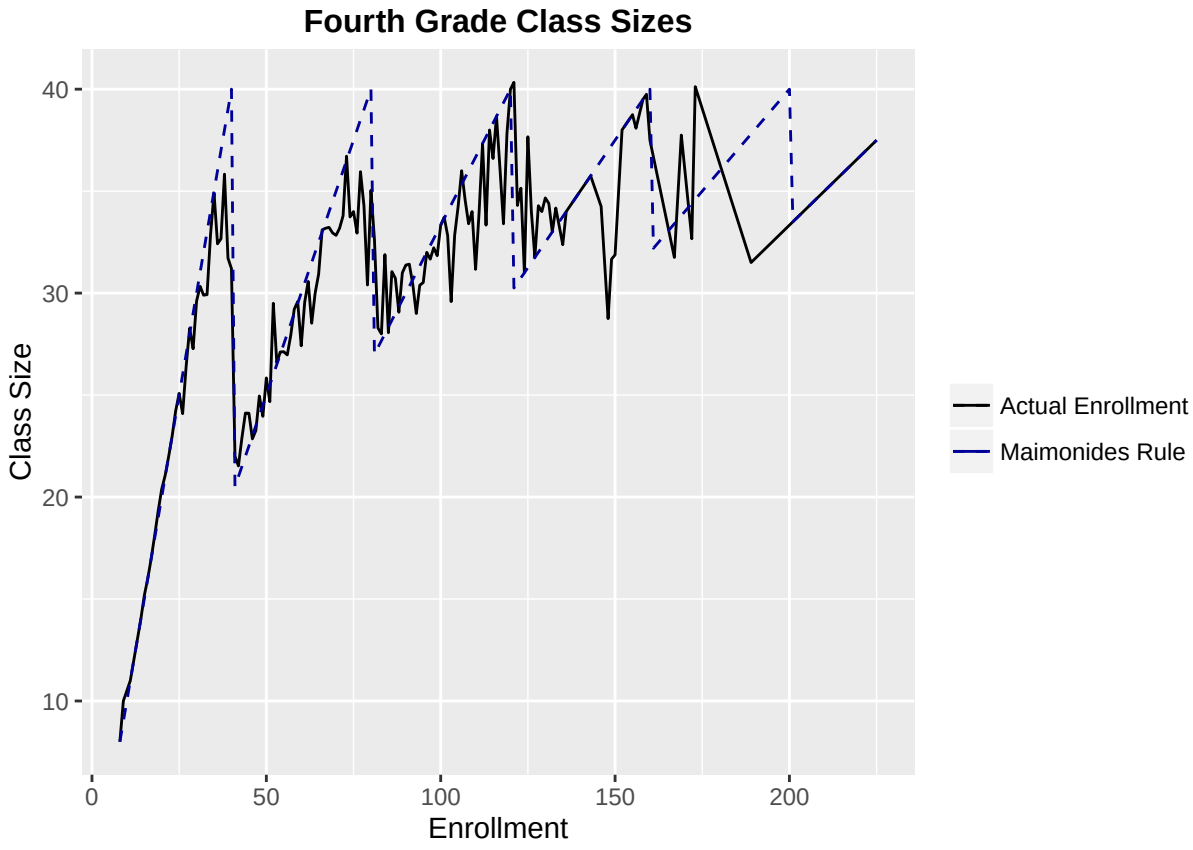


pdf

Table 5: OLS estimates for 1991 5th grade

	rc5.1	rc5.2	rc.3	m5.1	m5.2	m5.3
Class size	0.221	−0.031	−0.025	0.330	0.089	0.028
	(0.034)	(0.026)	(0.033)	(0.041)	(0.038)	(0.043)
Percent Disadvantaged		−0.350	−0.351		−0.335	−0.327
		(0.014)	(0.015)		(0.019)	(0.019)
Enrollment			−0.002			0.018
			(0.006)			(0.008)
R^2	0.035	0.367	0.367	0.050	0.240	0.243
Adj. R^2	0.035	0.366	0.366	0.049	0.239	0.242
Num. obs.	2019	2019	2019	2019	2019	2019
RMSE	7.548	6.116	6.118	9.470	8.471	8.456

rc := reading comprehension, m := math



2
pdf 2

TABLE II

OLS Estimates for 1991

This is a preliminary OLS estimate of some of the variables of interest in the data. It shows that by controlling for percent disadvantaged and total enrollment, any positive correlation between class size and test scores is eliminated, and all of the coeffecients on classize are negative. Thus suggesting that as class sizes increase, test scores decrease.

Table 6: OLS estimates for 1991 4th grade

	rc5.1	rc5.2	rc.3	m5.1	m5.2	m5.3
Class size	0.141	−0.053	−0.040	0.221	0.055	0.009
	(0.035)	(0.028)	(0.032)	(0.039)	(0.036)	(0.040)
Percent Disadvantaged		−0.339	−0.341		−0.289	−0.281
		(0.015)	(0.016)		(0.017)	(0.017)
Enrollment			−0.004			0.014
			(0.006)			(0.007)
R^2	0.013	0.309	0.309	0.025	0.204	0.207
Adj. R^2	0.012	0.309	0.308	0.025	0.204	0.206
Num. obs.	2049	2049	2049	2049	2049	2049
RMSE	7.943	6.645	6.646	8.658	7.825	7.815

rc := reading comprehension, m := math

Table 7: Reduced-Form Estimates for 1991 5th Grade Full Sample

	cs5.1	cs5.2	rc5.1	rc5.2	m5.1	m5.2
Expected Class Size	.704	.542	−.111	−.150	.006	−.113
	(.025)	(.037)	(.029)	(.039)	(.043)	(.053)
Percent Disadvantaged	−.076	−.053	−.359	−.354	−.349	−.332
	(.011)	(.010)	(.014)	(.015)	(.019)	(.019)
Enrollment		.043		.010		.032
		(.006)		(.006)		(.008)
R^2	.516	.553	.374	.375	.237	.246
Adj. R^2	.516	.552	.373	.374	.236	.245
Num. obs.	2019	2019	2019	2019	2019	2019
RMSE	4.555	4.380	6.085	6.079	8.489	8.441

rc := reading comprehension, m := math

TABLE III

Reduced-Form Estimates for 1991

Expected Class Size, Dependent on Enrollment

This table has two important improvements over the last table. We include a variable for expected class size, which is the class size that the Maimonides' Rule returns for a given enrollment level. Additionally we run the regressions on a discontinuous sample; the discontinous sample is made by only including the schools that have enrollment levels on the threshold of Maimonides' suggested class additions. This checks the sensitivity/accuracy of the estimates by focusing on the edge cases.

TABLE IV

2SLS Estimates for 1991 (5th Graders)

These models use the expected class size defined in the last model as an instrumental variable for class size, and defines a trend variable with slopes identical to the slope of the expected class size function's linear segments. This trend tests whether the discontinuity generating instrumental variable (expected class size) is adequately controlled for. The presence of strong negative coeffecients on class size in both the full and

Table 8: Reduced-Form Estimates for 1991 4th Grade Full Sample

	cs5.1	cs5.2	rc5.1	rc5.2	m5.1	m5.2
Expected Class Size	.772	.670	−.085	−.089	.038	−.033
	(.023)	(.033)	(.031)	(.040)	(.040)	(.050)
Percent Disadvantaged	−.054	−.039	−.340	−.340	−.292	−.282
	(.009)	(.009)	(.015)	(.016)	(.017)	(.017)
Enrollment		.027		.001		.019
		(.005)		(.007)		(.008)
R^2	.561	.575	.311	.311	.203	.207
Adj. R^2	.560	.574	.311	.310	.203	.206
Num. obs.	2049	2049	2049	2049	2049	2049
RMSE	4.195	4.129	6.635	6.637	7.829	7.813

rc := reading comprehension, m := math

Table 9: Reduced-Form Estimates for 1991 5th Grade Discontinuity Sample

	cs5.1	cs5.2	rc5.1	rc5.2	m5.1	m5.2
Expected Class Size	.481	.346	−.197	−.202	−.089	−.154
	(.057)	(.062)	(.050)	(.060)	(.072)	(.079)
Percent Disadvantaged	−.130	−.067	−.424	−.422	−.435	−.405
	(.033)	(.028)	(.036)	(.036)	(.040)	(.041)
Enrollment		.086		.003		.041
		(.016)		(.015)		(.020)
R^2	.360	.437	.421	.421	.296	.305
Adj. R^2	.357	.434	.419	.418	.293	.301
Num. obs.	471	471	471	471	471	471
RMSE	5.948	5.583	6.237	6.243	8.579	8.531

rc := reading comprehension, m := math

Table 10: Reduced-Form Estimates for 1991 4th Grade Discontinuity Sample

	cs5.1	cs5.2	rc5.1	rc5.2	m5.1	m5.2
Expected Class Size	.625	.503	−.061	−.075	.059	.012
	(.048)	(.061)	(.059)	(.064)	(.080)	(.080)
Percent Disadvantaged	−.068	−.029	−.348	−.343	−.306	−.291
	(.029)	(.027)	(.035)	(.038)	(.040)	(.042)
Enrollment		.063		.007		.024
		(.016)		(.016)		(.020)
R^2	.428	.475	.299	.299	.178	.182
Adj. R^2	.425	.471	.295	.294	.174	.176
Num. obs.	415	415	415	415	415	415
RMSE	5.485	5.261	6.574	6.580	8.259	8.246

rc := reading comprehension, m := math

Table 11: 2SLS Estimates for 1991 5th Grade Full Sample

	rc5.1	rc5.2	rc5.3	rc5.4	m5.1	m5.2	m5.3	m5.4
Class Size	−.158	−.277	−.263		.009	−.208	−.249	
	(.042)	(.076)	(.094)		(.061)	(.101)	(.124)	
Percent Disadvantaged	−.371	−.369	−.369		−.349	−.344	−.344	
	(.016)	(.016)	(.016)		(.021)	(.021)	(.021)	
Enrollment		.022	.013			.041	.067	
		(.009)	(.026)			(.012)	(.036)	
Enrollment squared/100			.004				−.012	
			(.010)				(.014)	
Trend				.137				.195
				(.036)				(.043)
R^2	.357	.340	.343	-.004	.237	.228	.223	.031
Adj. R^2	.356	.339	.341	-.005	.237	.227	.221	.030
Num. obs.	2019	2019	2019	1961	2019	2019	2019	1961
RMSE	6.166	6.249	6.236	7.720	8.486	8.540	8.569	9.592

rc := reading comprehension, m := math

Table 12: 2SLS Estimates for 1991 5th Grade Discontinuous Sample

	rc5.1	rc5.2	m5.1	m5.2
Class Size	−.410	−.582	−.185	−.443
	(.118)	(.206)	(.155)	(.251)
Percent Disadvantaged	−.477	−.461	−.459	−.435
	(.049)	(.047)	(.052)	(.050)
Enrollment		.053		.079
		(.032)		(.037)
R^2	.314	.240	.261	.209
Adj. R^2	.311	.235	.258	.204
Num. obs.	471	471	471	471
RMSE	6.791	7.154	8.787	9.101

rc := reading comprehension, m := math

discontinuous samples suggests causality (due to the IV and Discontinuous characteristics) and imply that there is causal relationship between class size increasing and test scores decreasing.

```
## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.

## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.

## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.

## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.
```

Table 13: 2SLS Estimates for 1991 4th Grade Full Sample

	rc5.1	rc5.2	rc5.3	rc5.4	m5.1	m5.2	m5.3	m5.4
Class Size	−.110	−.133	−.074		.049	−.050	−.033	
	(.040)	(.061)	(.068)		(.052)	(.075)	(.085)	
Percent Disadvantaged	−.346	−.345	−.346		−.290	−.284	−.284	
	(.016)	(.016)	(.016)		(.018)	(.017)	(.017)	
Enrollment		.005	−.040			.020	.007	
		(.008)	(.022)			(.009)	(.027)	
Enrollment squared/100			.021				.006	
			(.009)				(.012)	
Trend				.100				.130
				(.026)				(.029)
R^2	.307	.306	.312	.002	.204	.206	.206	.030
Adj. R^2	.307	.305	.311	.001	.204	.204	.205	.029
Num. obs.	2049	2049	2049	2001	2049	2049	2049	2001
RMSE	6.654	6.663	6.635	8.017	7.825	7.820	7.818	8.649

rc := reading comprehension, m := math

TABLE V

2SLS Estimates for 1991 (4th Graders)

This table applies the same models used in Table IV on 5th graders, to the data for 4th graders.

```
## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.

## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.

## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.

## Warning in override(models, override.coef, override.se, override.pval,
## override.ci.low, : Number of p values provided does not match number of
## models. Using default p values.
```

TABLES VI - VII

We did not include these tables due to time restrictions and strange results when using indicator instrumental variables with `ivreg()`. If you are interested in the code (we attempted to reproduce the results but could not finish these models accurately) please email ntadelis@berkeley.edu

CONCLUSION

This project is in our minds successful. We have gained a much deeper understand of R's statistical analysis tools, and the versatility of CRAN's compendium of packages. Additionally our results were within rounding

Table 14: 2SLS Estimates for 1991 4th Grade Discontinuous Sample

	rc5.1	rc5.2	m5.1	m5.2
Class Size	−.098	−.150	.095	.023
	(.095)	(.132)	(.095)	(.132)
Percent Disadvantaged	−.354	−.347	−.299	−.290
	(.036)	(.038)	(.036)	(.038)
Enrollment		.017		.023
		(.022)		(.022)
R^2	.284	.275	.183	.184
Adj. R^2	.281	.270	.179	.178
Num. obs.	415	415	415	415
RMSE	6.642	6.690	8.229	8.239

rc := reading comprehension, m := math

errors of Angrist and Lavy's regression results, suggesting that any differences between Stata and R are so small that when regression models are run correctly they are all but eliminated.