Effects of 2003 and 2011 Residency Work-Hour Reforms on Exam Pass Rates

Olivia Fu, Brian Kim, Sophia Yang

September 03, 2025

1 Background and Motivation

Medical residency is a demanding stage of training with long work hours, so in 2003 and 2011 reforms were added that limited the number of hours medical residents could work. In 2003, residents became limited to a maximum of 80 hours per week while in the 2011 reform placed more specific limitations on the types of activities. We want to determine if changes in work hours affected residency exam pass rates. In particular, did more study time from the limited work hours lead to higher pass rates or did it have the opposite effect by providing them less clinical exposure and therefore leading to a lower pass rate? The goal of our analysis is to determine how exam pass rates do or do not differ across the three time periods of pre-reform (1996–2002), 2003 reform (2003–2010), and 2011 reform (2011–2015).

2 Data and Exploratory Analysis

Our dataset contains data for first-time medical residency examinees in an undisclosed field. Each row represents a single year. It has three variables representing the year (`Year`), number of students taking the exam (`N`), and the percentage of examinees who passed (`Pct`).

To begin, we plotted certification exam pass rates and number of examinees by year in Figure 1. In this plot, we observed a sharp increase in pass rates after the 2003 reform with pass rates peaking in 2007, before declining. The decline in pass rates seems to coincide with an increase in the number of students attempting the exam from 2007 to 2010, raising into question if this change in pass rate may partially be the result of less qualified students who previously avoided the exam choosing to take the resident exam. Yet, despite continual increases in the number of examinees, pass rates eventually rose again starting in 2011. While the percentage of passes in 2015 is less than the peak in 2007, it is still higher than pass rates in the pre-reform period.

Table 1 shows mean pass rates and slopes by reform period. The pass rates were lowest on average during pre-reform period with a mean of about 85.3% but also showed an average increase in pass rate by 0.821% per year. During the 2003 reform the mean pass rate increased to 90.9% while pass rates trended downward, declining by 0.655% each year. Although the pass rates were on average higher than during other time periods, performance declined during this period. Finally, during the 2011 reform period, the mean pass rates returned to near pre-reform rates at about 86.2%. However, the positive slope of 1.2% could suggest a higher average exam pass rate in the future.

3 Model Selection, Implementation, and Evaluations

3.1 Selection

We chose a quasibinomial model with a logit link to analyze the pass rates of medical residents. For the sampling model, each year's outcome is the number of residents passing out of the total number of test-

takers. This naturally suggests a binomial model, since each candidate either passes or fails in a given year. However, analysis in class revealed evidence of overdispersion in this dataset, so we used the quasibinomial specification to allow the variance to be larger than the standard binomial assumption.

- $S_y \sim \text{Binomial}(N_y, \pi_y)$, where S_y = number passing in year y.
- N_y = number of test-takers.
- π_y = probability of passing in year y.

To model the passing rate, we employed a logistic regression. The logit link ensures that predicted probabilities stay between 0 and 1. Intuitively, the model is piecewise linear in the log-odds of passing, with separate linear trends estimated for the three time periods.

$$\text{logit}(\pi_y) \;=\; \beta_0 \;+\; \beta_1 \cdot \text{Year} \;+\; \beta_2 \cdot 1_{\{tp2\}} \;+\; \beta_3 \cdot 1_{\{tp3\}} \;+\; \beta_4 \cdot (\text{Year} \times 1_{\{tp2\}}) \;+\; \beta_5 \cdot (\text{Year} \times 1_{\{tp3\}})$$

The choice of predictors was motivated by exploratory plots (Figure 1) and summary tables (Table 1), which showed distinct level shifts and slope changes across periods separated by the 2003 and 2011 reforms. We therefore included indicators for time periods and their interactions with Year. This formulation allows each time period to have its own intercept and slope, thereby capturing both immediate reform effects (intercept shifts) and longer-term changes in trend (slope differences).

We also considered the effect of cohort size. Figure 1 shows that the number of test-takers varied each year, with a gradual increase from 2003 to 2015. To evaluate the potential impact of cohort size on pass rates, we also added N (number of test-takers) to the model. Since the effect was not statistically significant, we did not include N in the final model.

3.2 Implementation

Models were fit in R using the `glm()` function with a quasibinomial family and logit link. The response was specified as `cbind(Pass, Fail)` to reflect binomial counts. `Year` was centered at the relevant reform year (2003 or 2011) to make the intercepts more interpretable. Interaction terms between Year and time periods were included to estimate both intercept shifts and slope changes associated with the reforms.

3.3 Evaluation

We evaluated the model using both tests and diagnostic plots.

We first conducted a drop-in-deviance test to evaluate whether including interaction terms improved the model. As shown in Table 2, the deviance dropped substantially when moving from the simpler additive model to the model with interactions, and the associated F-test had a very small p-value. This provides strong evidence that we should keep the interaction terms between Year and time period, allowing each time period to have its own slope.

Second, we assessed overdispersion. The estimated dispersion parameter (hat =12.24) is much larger than 1, confirming substantial overdispersion. Likely contributors include: (i) outliers in the dataset, (ii) heterogeneity in test-takers' passing probabilities, and (iii) unobserved factors such as exam difficulty or changes in scoring standards across years. The quasibinomial model accounts for the extra-variation in the data by adjusting the estimated variance.

Finally, we assessed goodness-of-fit using deviance residual plots. Figure 2 (residuals vs. fitted values) shows residuals mostly scattered around 0 with no clear patterns, indicating the model fits reasonably well. One notable exception is 2007, which has a very large residual; with only 20 observations in our dataset, this outlier is influential and likely inflates overdispersion. We also examined residuals vs. Year (Figure 3) to assess whether the model reflected trends over time. Residuals again scatter randomly around 0 with no systematic trends, suggesting the piecewise-linear structure (tp1/tp2/tp3 with separate slopes) is adequate to capture time patterns.

4 Results

Our model found that the reform in 2003 (`timeperiodtp2`) had a positive increase in log-odds pass rates by 0.571 relative to the pre-reform time period (Table 3). This implies that without the 2003 reform, the log-odds of passing in 2003 would have been 0.571 lower than observed. Using an alpha of 0.05, this result is statistically significant with a p-value of 0.001. This is contrary to the negative slope of the period, potentially indicating a short term positive effect (Figure 4). This could be a signal that in the long term the effects are diminished as students get used to the new normal.

In contrast, the 2011 reform (`timeperiodtp3`) had a slight decrease in the log-odds of passing by 0.313 relative to the preceding period (Table 4). While slight, this decrease is significant with a p-value of 0.039 and the 95% confidence interval does not include 0 (-0.583, -0.045). This result was derived by fitting the same model, but with `Year` recentered at 2011 and time period 2 set as the baseline. While pass rates were overall lower than they were in time period 2, they revert the prior trend of declining pass rates and appear to continually trend upwards (Figure 4). Given more time, the 2011 reform may be a long term beneficial change.

5 Limitations and Conclusion

Our analysis has certain limitations. Due to the limited years of data, it is difficult to establish a strong baseline and effect of the model. Our model uses a single dispersion parameter and hence assumes overdispersion is constant rather than varying over time. We also did not take into account the length of the typical residency program of 3 years. This length might indicate the presence of a lagged effect in which the full effect of the reform is not felt until two or three years after it began. This could be addressed by weighing the year further after a reform more or by using lagged versions of variables (e.g. considering the second year after a reform as the actual first year of the reform). Additional directions for improvement include introducing more variability as exam takers are likely a heterogeneous group. Referencing of other adjacent datasets such as from the MCAT could also help us evaluate if there were changes in the caliber of student entering residency programs.

In summary, we found that there were significant changes in residency pass rates following both reforms. From our model, we observed that the reform in 2003 had a significant but potentially temporary positive effect on pass rates for the medical residency exam. On the other hand, when compared to the 2003 reform period, the reform in 2011 had a slight negative effect. However, the 2011 reform appears to have reversed a downward trend in pass rates and may have a long lasting positive effect in the future.

6 Appendix

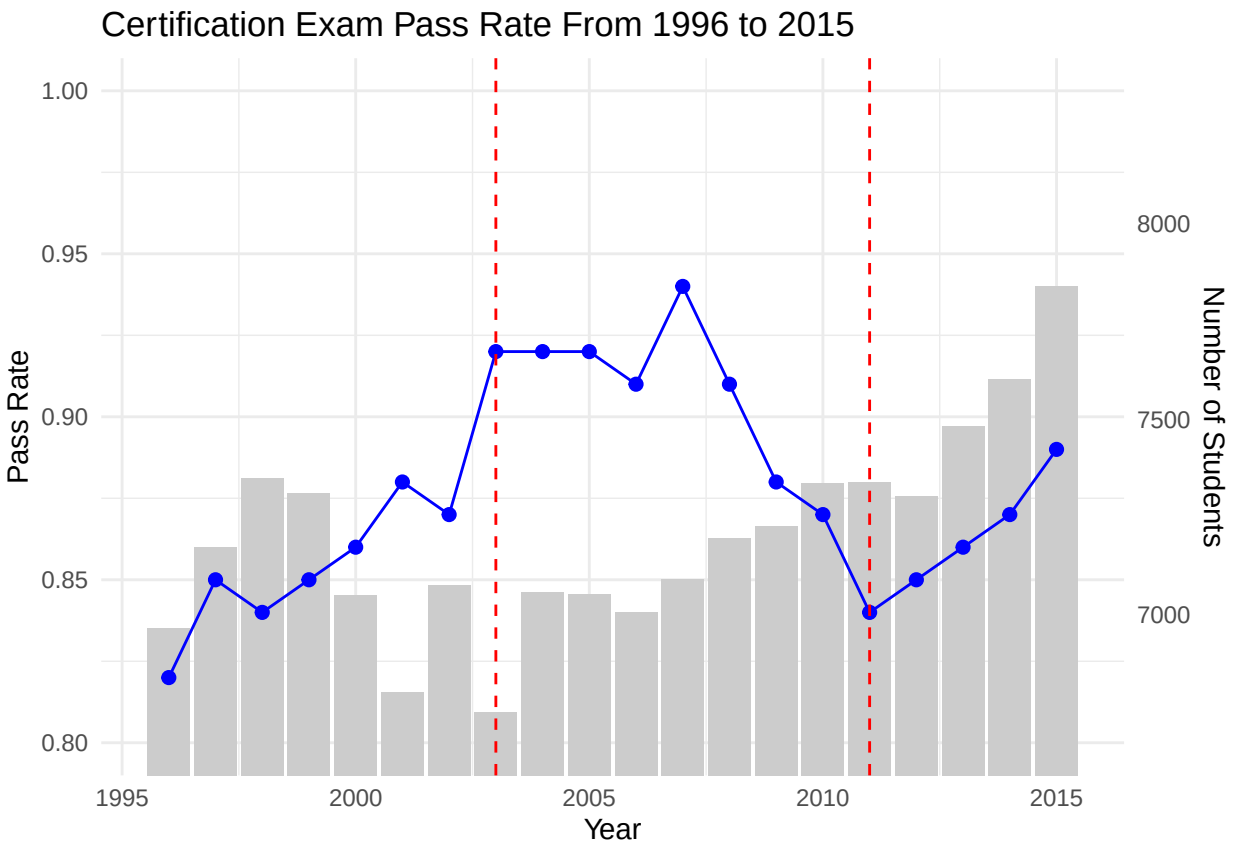

Figure 1: Blue points represent the proportion of test takers who passed. Gray bars represent the total number of students taking the exam, labeled on the right axis. Observe that the increase in pass rate from 2000 to 2003 aligns with a decrease in the number of students taking the exam. Similarly, pass rates declined in 2007-2011 while the number of students increased over the same time frame.

Table 1: Mean pass rates and slope by period.

timeperiod	MeanPassRate	SlopePct
Pre-reform	0.853	0.821
2003 Reform	0.909	-0.655
2011 Reform	0.862	1.200

Table 2: Drop-in-deviance comparing additive vs interaction quasibinomial models.

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
16	534.199	NA	NA	NA	NA
14	183.044	2	351.155	14.344	0

Table 3: Coefficients with Year centered at 2003 and time period 1 as the baseline (quasibinomial logit).

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.024	0.104	19.488	0.000	1.823	2.230
Year_centered_2003	0.065	0.022	2.916	0.011	0.021	0.109
timeperiodtp2	0.571	0.145	3.938	0.001	0.287	0.856
timeperiodtp3	-1.201	0.385	-3.121	0.008	-1.954	-0.445
Year_centered_2003:timeperiodtp2	-0.146	0.032	-4.598	0.000	-0.208	-0.084
Year_centered_2003:timeperiodtp3	0.036	0.043	0.842	0.414	-0.048	0.122

Table 4: Coefficients with Year centered at 2011 and time period 2 as the baseline (quasibinomial logit).

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.950	0.106	18.464	0.000	1.745	2.159
Year_centered_2011	-0.081	0.022	-3.585	0.003	-0.125	-0.037
timeperiodtp1	0.596	0.296	2.014	0.064	0.019	1.179
timeperiodtp3	-0.313	0.137	-2.283	0.039	-0.583	-0.045
Year_centered_2011:timeperiodtp1	0.146	0.032	4.598	0.000	0.084	0.208
Year_centered_2011:timeperiodtp3	0.182	0.043	4.204	0.001	0.098	0.268

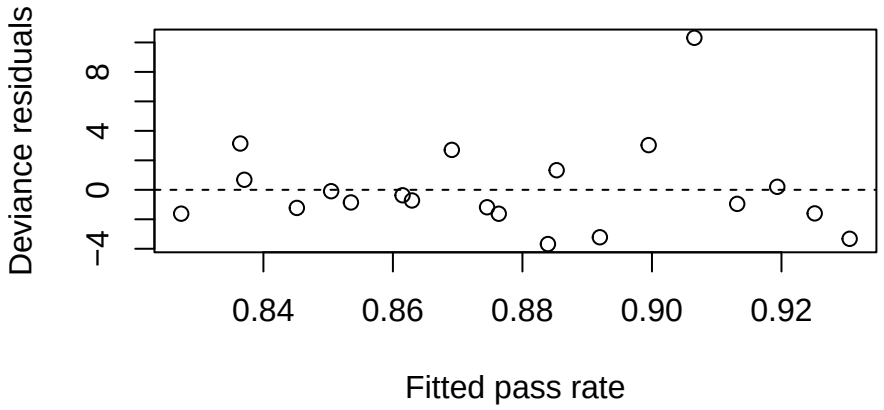

Figure 2: Deviance residuals vs. fitted pass rate. Points cluster around the zero line with no clear pattern, indicating an adequate overall fit. One large residual in 2007 stands out; with only 20 observations, this outlier is influential and likely inflates overdispersion.

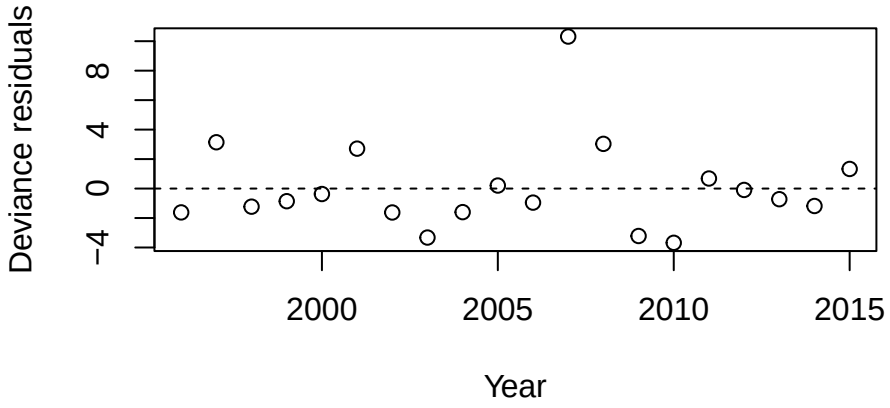

Figure 3: Deviance residuals vs. year. Residuals remain randomly scattered around zero without visible time trends, supporting the piecewise-linear specification (tp1/tp2/tp3 with separate slopes) as properly capturing temporal patterns.

Figure 4: Lines show the predicted pass rate for that year based on the quasibinomial model. The shaded areas represent the 95% confidence intervals. Black points represent actual observations. Note positive trends in pass fail rate during the 1996-2002 and 2011-2015 time periods. 2003-2010 shows an initial increase in pass rates but then declines in the later two years.