1 Problem 1

1.1 Question 1

Table 1 shows summary statistics for the entire sample and for firms with a degree of unionization above/below 50 percent.

The data set includes observations on 36,495 workers. There are 1,000 firms and about 85 workers per firm on average. The median is 60 workers per firm indicating that the firm size distribution is skewed. Indeed, the largest firm in the sample has 394 workers. More than half of the workers in the sample (21,132/36,495) are potentially covered by union bargaining. The median wage level is 276,000 DKK, but the dispersion is large with some earning more than 600,000 DKK and others earnings only 100,000 DKK.

Based on the median, the level of earnings appears to be slightly higher in firms potentially covered by union wage bargaining. These firms also appear to be slightly bigger than firms with a unionization rate below 50. However, the firm size difference does not come out so clearly when judged by the mean. The fraction of women and the age distribution appear to be similar across firms covered by unionized wage bargaining and not.

In summary, the descriptive statistics indicate that there is a potential for a unionized wage bargaining premium to exist.

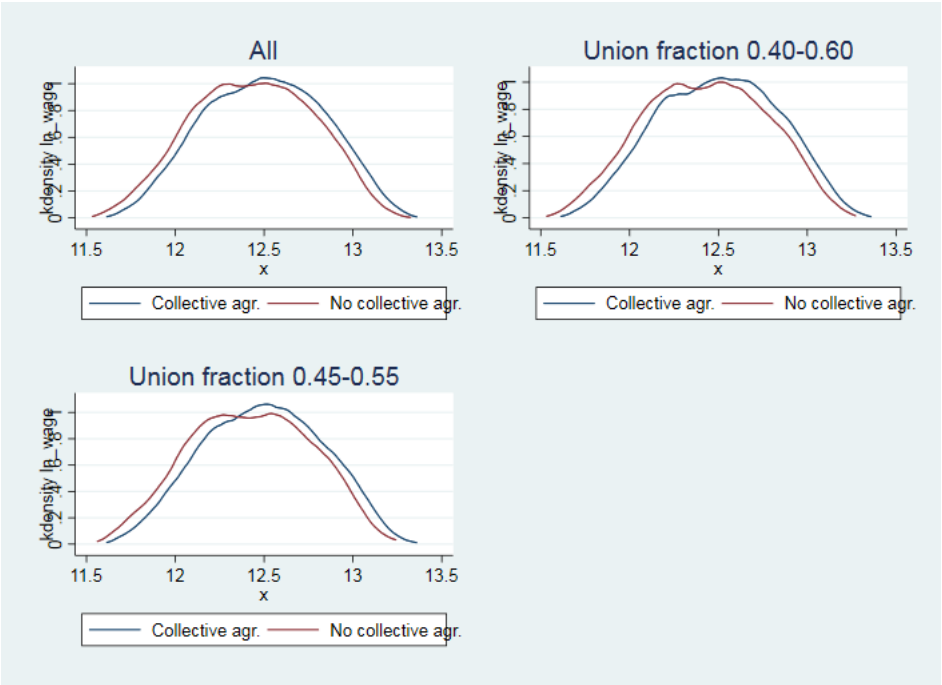
Union frac>=50	variable	N	mean	p50	min	max
0	number of employees	15363	86.4725	57	0	394
	union	15363	.3826727	0	0	1
	age	15363	39.38918	39	20	59
	female	15363	.6480505	1	0	1
	wage	15363	265586	251921.9	102048.1	611561.9
1	number of employees	21132	84.38515	63	0	394
	union	21132	.6059057	1	0	1
	age	21132	39.5229	40	20	59
	female	21132	.6539372	1	0	1
	wage	21132	283993	269392.3	110550.9	634096.1
Total	number of employees	36495	85.26384	60	0	394
	union	36495	.5119331	1	0	1
	age	36495	39.46661	40	20	59
	female	36495	.6514591	1	0	1
	wage	36495	276244.3	262051.3	102048.1	634096.1

1.2 Question 2

Below, we show the kernel densities of log wages.[1] In the first graph all observations are used, whereas in the second and third graphs only include observations for firms with union coverage of, respectively, $0.40 - 0.60$ and $0.45 - 0.55$. All graphs broadly show the same picture, that wage with union agreements

[1] It is ok if wage levels are used in stead of log wages although the latter usually is most appropriate.

1.2 Density of log wages, all data



stochastically dominates those without union agreements. This is in accordance with the trade union theory.

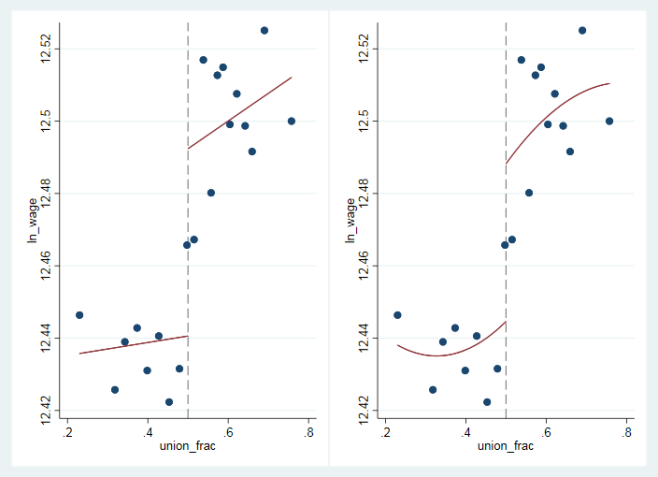
2 Problem 2

2.1 Question 1

We can use RD to estimate the effect of being covered by a wage agreement when there is a jump in the probability of being in covered by a union. In the question it does not completely rule out that union wage agreements could exist for firms with less than 50 percent union members. We assume that this is not the case and, hence, that we have a sharp RD, where the probability of being covered by a union increases discretely from 0 to 1 at the 50 percent cutoff. The RD design boils down to comparing wages just below and above the 50 percent cutoff and attributing the difference in the wage below and above to the union wage agreement. We only estimate a local treatment effect, which holds at the 50 percent cutoff.

However, we can only use RD if there is no manipulation of the running variable. Therefore, if it is the case that workers influence each other to join a union when for example the profits are large, the RD design will be invalid for estimating the effect of being covered by a wage agreement.

2.2 Regression discontinuity with a linear specification



2.2 Question 2

We have shown two binned scatterplots. In the first, we have a linear fit on each side of the cutoff and in the second a quadratic fit. Both show a clear jump in the wages at the cutoff of 50 percent. From eye-balling, the preliminary effect seems to be 7 percent.

2.3 Question 3

2.3.1 (a)

We are asked to estimate a linear model with the same slope on each slides without using control variables. This amounts to estimating the following regression model

$$\ln w_i = \beta_0 + \beta_1 Ufrac + \rho \left(Ufrac \geq 0.50\right) + u_i$$

where i denotes the individual worker and u_i is an error-term. We are primarily interested in estimating the treatment effect ρ. The parameter estimates is shown in the table below. In the first column, we use heteroscedasticity robust standard errors, whereas we in the second use standard errors clustered on firm level. It seems reasonable to cluster on firm level since workers in the same firm may be subject to similar shocks. Furthermore, the experimental variation is at the firm-level. In both cases, our treatment effect estimate at 6.9 percent is significant at a 1 percent level, but the clustered standard errors are larger. Hence, we will use clustered standard errors in the estimations to follow.

Simple model with same slopes on both sides and no control variables		
	Robust standard errors	Clustered standard errors (on firm)
ρ	0.069***	0.069***
	(0.006)	(0.009)
β_1	-0.004	-0.004
	(0.023)	(0.032)
β_0	12.435***	12.435***
	(0.009)	(0.013)

* p<0.10, ** p<0.05, *** p<0.01

2.4 (b)

Compared to the previous model, we now include control variables while still restricting the slope to be the same on both sides of the cutoff. Including covariates does not change the estimate of the treatment effect, nor its significance. We see that older workers and men, on average, have higher earnings.

Simple model with same slopes on both sides and control variables	
ρ	0.070***
	(0.009)
β_1	-0.010
	(0.032)
age	0.003***
	(0.000)
female	-0.071***
	(0.004)
num_emp	-0.000
	(0.000)
β_0	12.370***
	(0.014)

* p<0.10, ** p<0.05, *** p<0.01

2.4.1 (c)

We asked to estimate a model, which allows for different slopes on each side of the cutoff. It is not specified in the question whether we should include control variables, so we estimate the model with and without control variables. Let $D_i = (Ufrac_i \geq 0.50)$, then we can write the model as

$$\begin{aligned} \ln w_i &= \beta_0 + \beta_1 (Ufrac_i - 0.50) + \beta_2 (Ufrac_i - 0.50) \cdot D_i + \rho D_i \\ &\quad + \beta_3 age_i + \beta_4 female_i + \beta_5 num\_emp_i + u_i \end{aligned}$$

The two specifications imply estimates of 7.0 and 7.1 percent higher wage with a collective agreement. Both estimates are significant at 1 percent, but not significantly different. The main conclusion is that allowing for different slopes

only changes the estimate marginally. The effect of the control variables is similar to problem 2, question 3 (b).

Model with different linear slopes on each side of the cutoff		
	Without control variables	With control variables
ρ	0.070***	0.071***
	(0.009)	(0.009)
β_1	-0.077	-0.075
	(0.048)	(0.048)
β_2	0.140**	0.124*
	(0.064)	(0.063)
age (β_3)		0.003***
		(0.000)
female (β_4)		-0.071***
		(0.004)
num_emp (β_5)		-0.000
		(0.000)
β_0	12.424***	12.356***
	(0.008)	(0.011)

* p<0.10, ** p<0.05, *** p<0.01

2.4.2 (d)

In this question, we are going to allow for quadratic trends on each side of the 50 percent cutoff. We state the model without control variables

$$\begin{aligned} \ln w_i &= \beta_0 + \beta_1 (Ufrac_i - 0.50) + \beta_2 (Ufrac_i - 0.50) \cdot D_i \\ &\quad + \beta_3 (Ufrac_i - 0.50)^2 + \beta_4 (Ufrac_i - 0.50)^2 \cdot D_i + \rho D_i + u_i \end{aligned}$$

However, we estimate the model with and without control variables. Allowing for quadratic trends implies that the parameter estimate increases slightly to 7.5 and 7.6 percent depending on whether control variables are included. Again, we obtain similar estimates for the control variables as in the previous questions.

Model with different quadratic slopes on each side of the cutoff		
	Without control variables	With control variables
ρ	0.075***	0.076***
	(0.012)	(0.011)
β_1	-0.205	-0.210*
	(0.128)	(0.124)
β_2	0.302*	0.306*
	(0.163)	(0.159)
β_3	-0.415	-0.439
	(0.361)	(0.356)
β_4	0.293	0.268
	(0.479)	(0.480)
age		0.003***
		(0.000)
female		-0.071***
		(0.004)
num_emp		-0.000
		(0.000)
β_0	12.418***	12.350***
	(0.009)	(0.012)

* p<0.10, ** p<0.05, *** p<0.01

2.4.3 (e)

In order to estimate heterogeneous effects of a variable x, we interact the indicator function for being above the cutoff, i.e. $D_i = (Ufrac_i \geq 0.50)$, with the variable while subtracting this variable's mean. Hence, for each x variable we will include terms such as $D_i (x_i - \bar{x})$, where $\bar{x} = N^{-1} \sum_{i=1}^{N} x_i$. Subtracting the mean of the variables makes sure that we still capture the overall treatment effect by the parameter ρ in the following equation

$$\begin{aligned} \ln w_i \quad = \quad & \beta_0 + \beta_1 D_i \left(female_i - \overline{female} \right) + \beta_2 D_i \left(age_i - \overline{age} \right) + \beta_3 \left(Ufrac_i - 0.50 \right) \\ & + \beta_4 \left(Ufrac_i - 0.50 \right) \cdot D_i + \beta_5 \left(Ufrac_i - 0.50 \right)^2 + \beta_6 \left(Ufrac_i - 0.50 \right)^2 \cdot D_i \\ & + \rho D_i + \beta_7 age_i + \beta_8 female_i + \beta_9 emp\_num_i + u_i \end{aligned}$$

Model with heterogenous effects	
ρ	0.076***
	(0.011)
β_1 (effect of female)	0.050***
	(0.007)
β_2 (effect of age)	0.002***
	(0.000)
β_3	-0.204
	(0.125)
β_4	0.300*
	(0.159)
β_5	-0.436
	(0.358)
β_6	0.260
	(0.481)
age (β_7)	0.002***
	(0.000)
female (β_8)	-0.100***
	(0.005)
num_emp (β_9)	-0.000
	(0.000)
β_0	12.407***
	(0.015)
* p<0.10, ** p<0.05, *** p<0.01	

The results show that whereas women, in general, receive lower wages (β_8) they benefit more from being in union wage agreements than men (β_1). We see that older workers, in general, earn more than younger workers probably due to more experience (β_7). In addition to this, older workers benefit more from being in a union wage agreement (β_2).

The regression involves sample averages computed prior to the estimation. To obtain the right standard errors, we should bootstrap over the calculation of means and the regression. However, in practice this is a minor concern since sample means are precisely estimated. Thus, we have just computed clustered standard errors clustered at the firm-level.

2.4.4 (f)

The earnings difference between a 50 years old woman covered by a union agreement and a 30 years old man not covered by a union is

$$\rho \cdot 1 + \beta_1 \left(1 - \overline{female}\right) + \beta_2 \left(50 - \overline{age}\right) + \beta_7 \left(50 - 30\right) + \beta_8 \cdot 1 =$$

$$0.076 + 0.050 \cdot (1 - 0.6515) + 0.002 \cdot (50 - 39.47) + 0.002 \cdot (50 - 30) - 0.100 = 0.051$$

We find that on average, the 50 years old woman covered by a union agreement earns 5 percent more. The standard errors of the estimated difference is 0.013, so the difference is significant at a 1 percent level.

3 Problem 3

3.1 Question 1

To minimize the risk of mistaking a discontinuity from a non-linearity, we can estimate the RD model nonparametrically. We will use a local linear nonparametric regression due to its better boundary properties compared to the local constant estimator. In practice, we simply estimate the following regression using only observations in a short interval around the cutoff and weighting each observation with the triangular kernel

$$\ln w_i = \beta_0 + \rho D_i + \beta_1 \left(Ufrac_i - 0.50\right) + \beta_2 \left(Ufrac_i - 0.50\right) \cdot D_i + u_i$$

Nonparametric RD			
Bandwidth	0.025	0.050	0.100
ρ	0.121***	0.084***	0.073***
	(0.017)	(0.019)	(0.015)
β_1	-3.212**	-0.729	-0.280
	(1.405)	(0.644)	(0.283)
β_2	0.517	0.641	0.458
	(1.779)	(0.777)	(0.330)
β_0	12.380***	12.408***	12.415***
	(0.014)	(0.016)	(0.013)

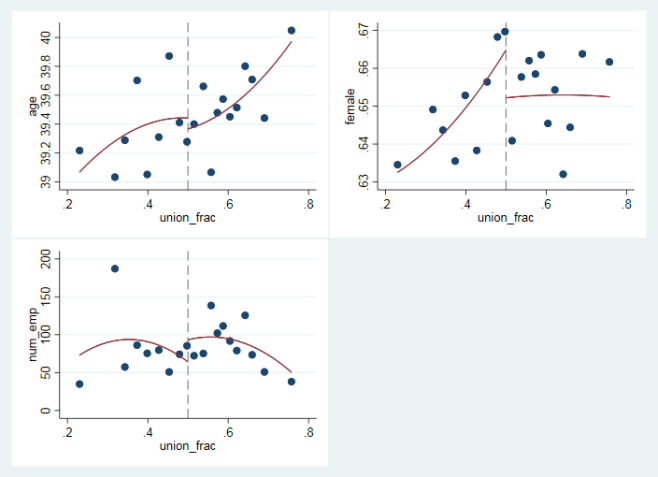
* p<0.10, ** p<0.05, *** p<0.01

We obtain a similar estimate of the treatment effect compared with the parametric specifications in the previous questions using a bandwidth of 0.100. However, with smaller bandwidths, the estimates are larger and even 0.121 for a bandwidth of 0.025. The estimate for each specific bandwidth is significant at a 1 percent level.

3.2 Question 2

In this question, we want to check that there is no discontinuity in the explanatory variables. Even though we control for explanatory variables, a (significant) discrete jump in any of the explanatory variables would require different specifications compared to what we have previously used. Besides this, it seems difficult to rationalize discontinuities in the control variables apart from perhaps the number of employees if recruitment is easier or labor turnover lower for firms with union bargained wages.

Regression discontinuity with age as dependent variable



First, we consider the question graphically by using binned scatterplots. Although the regression lines are not continuous through the 50 percentage coverage point, simple graphical inspection suggests that there is not a significant difference at the cutoff.

For the regressions, we will for each covariate k estimate the simplest specification with the same slope on both sides of the cutoff

$$covariate_{ki} = \beta_0 + \beta_1 D_i + \beta_2 Ufrac + u_i$$

and the most general parametric model allowing for quadratic functions of the running variable specific to each side of the cutoff

$$\begin{aligned} covariate_{ki} \quad = \quad & \beta_0 + \beta_1 D_i + \beta_2 \left(Ufrac_i - 0.50\right) + \beta_3 \left(Ufrac_i - 0.50\right) \cdot D_i \\ & + \beta_4 \left(Ufrac_i - 0.50\right)^2 + \beta_5 \left(Ufrac_i - 0.50\right)^2 \cdot D_i + u_i \end{aligned}$$

In both cases, β_1 captures the effect at the cutoff and for the design to be credible, we need that our β_1 estimate is insignificant. As show in the table below, β_1 is insignificant at all conventional significance levels for the two specifications for female and the number of employees. However, for age the effect at the cutoff is significant at a 5 percent level for the simple specification and at 10 percent for the more general model. Therefore, we also estimate the nonparametric RD specification for age using the same bandwidths as in the previous question. We see that the estimates are significant at a 5 percent level with bandwidths of 0.050 and 0.100, but insignificant with a bandwidth of 0.025. If we have enough observations close to the cutoff, we would with RD prefer working with a low bandwidth. Since we got significant effect with the log wage with a bandwidth of 0.025, there should be enough observations close to

the cutoff. Therefore, we conclude that there is no true discontinuity in age and that the significant results arise due to extrapolation away from the cutoff.

Estimating the effect of the cutoff on the control variables						
	age		female		num_emp	
β_1	-0.473**	-0.460*	-0.002	-0.003	2.446	-6.577
	(0.21)	(0.27)	(0.01)	(0.01)	(20.47)	(32.54)
β_2 (simple)	2.720***		0.036		-20.309	
	(0.75)		(0.03)		(69.86)	
β_2 (general)		2.666		0.134		-218.611
		(2.88)		(0.11)		(443.03)
β_3		-0.106		-0.200		594.749
		(3.91)		(0.15)		(461.05)
β_4		0.846		0.104		-938.295
		(9.02)		(0.35)		(1027.65)
β_5		0.809		0.048		-821.136
		(12.67)		(0.48)		(1091.81)
β_0	38.348***	39.685***	0.634***	0.662***	94.244***	79.843***
	(0.30)	(0.21)	(0.01)	(0.01)	(31.47)	(30.43)

p<0.10, ** p<0.05, *** p<0.01

Estimating the effect of the cutoff on age using nonparametric RD			
Bandwidth	0.025	0.050	0.100
ρ	-0.521	-0.807**	-0.663**
	(0.400)	(0.408)	(0.328)
β_1	-31.539	5.828	5.528
	(24.691)	(15.409)	(4.973)
β_2	74.289*	17.168	-0.698
	(37.845)	(18.860)	(6.529)
β_0	39.421***	39.795***	39.810***
	(0.304)	(0.330)	(0.252)

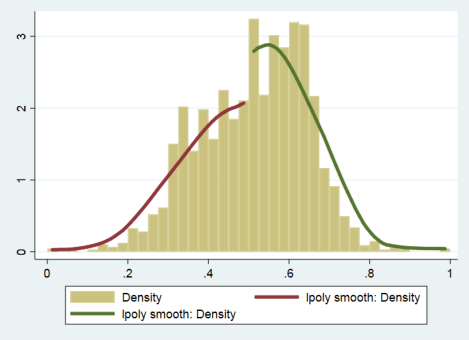
* p<0.10, ** p<0.05, *** p<0.01

3.3 Question 3

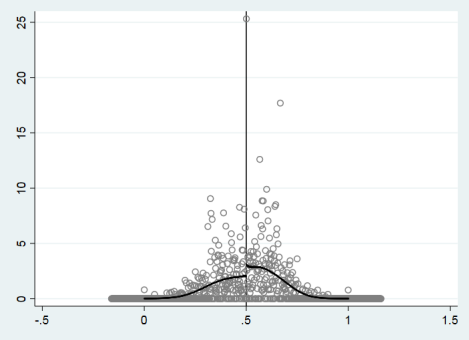
It is very important to examine whether there is manipulation of the running variable. In the current setting with union bargaining one could fear that union-covered workers could pressure each non-union members to become union members in order to get the union to bargain for them. Then, we would potentially have endogenous selection into union bargained wage agreements and this would violate our identifying assumptions. It would clearly be problematic if workers are more likely to join the union if the firm has market power and the workers want to get a bigger share of the profits.

The figures show the McCrary density test. In both cases a histogram is estimated by binning the data. Subsequently, a local linear regression is fit

Density test for manipulation of the running variable



Density test for manipulation of the running variable, DCdensity



on both sides of the cutoff. The figure clearly shows that the density is not the same on both sides of the 50 percent cutoff. McCrary's formal density test using `DCdensity` gives a discontinuity estimate of 0.401 and standard errors of 0.026. Hence, the null hypothesis of no bunching is strongly rejected. Therefore, there is manipulation of the running variable and the identifying assumptions of regression discontinuity design is not met. Therefore, the preceding analysis will not identify a causal effect of union bargaining.

4 Question 4

We know whether firms are covered in 2014 since this is a deterministic function of the fraction of union members in the firm in 2014. To avoid the problem of possible bunching at 35 percent in the union fraction in 2014, we could use that in 2013 there is no bunching at 35 percent. Hence, we can exploit this in a fuzzy regression discontinuity design by using a dummy for the union fraction being greater than 35 percent in 2013 as instrument for having a union wage agreement in 2014. More formally, let $\ln w_i^{2014}$ denote the log of the wage in 2014, let x_i^{2013} indicate the union fraction in 2013 for the firm the worker is em-

ployed by in 2014, D_i^{2014} indicate whether the individual works in a firm with a union wage agreement in 2014, and let T_i^{2013} denote a dummy for whether the union fraction in 2013 is greater than 35 percent. For the simplest case where we only assume linearity in the running variable, we have the model

$$\ln w_i^{2014} = \beta_0 + \beta_1 x_i^{2013} + \rho D_i^{2014} + \eta_i$$

with the following first stage equation

$$D_i^{2014} = \gamma_{00} + \gamma_1 x_i^{2013} + \pi T_i^{2013} + \xi_i$$

The estimate for ρ will capture the effect of being covered by a wage agreement and we will identify this using the local variation around 35 percent for the union fractions in 2013. It is important that we allow for polynomials in the specification such that we do not identify our estimate for ρ based on assuming linearity in the running variable x_i^{2013}. We can also estimate this instrumental variable model by local linear regression by weighting each observation around the 35 percent cutoff.

An alternative empirical strategy is to use diff-in-diff. We can use that workers employed in firms with union fractions between 35 and 50 percent will have a union wage agreement in 2014,but not in 2013. As control group, it seems most natural to use workers in firms with union fractions below 35 percent.

An alternative control group would be workers in firms with a union fraction of more than 50 percent sincethey are covered in both years. However, if dynamic effects of union wage agreements exist, this would not be an ideal control group. For example, suppose a firm was not covered before 2013 and that the full effect of union wage agreements only materializes after a couple of years. Then, using this firm as a control is not ideal since part of the change in wages between 2013 and 2014 is due to the firm only recently got a union wage agreement.

Let D_{it}^{2014} be a dummy for 2014, D_{it}^{35-50} be a dummy for being in a firm with a union fraction being in between 35 and 50 percent in 2013. Then, we would estimate the following model

$$\ln w_{it} = \beta_0 + \beta_1 D_{it}^{2014} + \beta_2 D_{it}^{35-50} + \rho D_{it}^{2014} \times D_{it}^{35-50} + \eta_{it}$$

The effect of a union wage agreement will be captured by ρ. One problem with the diff-in-diff specification is that workers can change firm and that the union fraction in the firm can change such that a firm with, say, a union fraction of 37 percent in 2013 only has a union fraction of 32 percent in 2014.

The diff-in-dif specification relies on an assumption of parallel trends. With only two observations we cannot assess the validity of this assumption.

For both types of identification strategies in this question, we only obtain the first year effect. This is unlike the sharp regression discontinuity design in the previous questions. Furthermore, with only two years of data we cannot examine whether dynamic effects are present.